White Paper **8/12/07**

Pathogenomics of Trypanosomatid parasites

Greg Buck (Virginia Commonwealth University)
Matt Berriman (Wellcome Trust Sanger Institute)
John Donelson (University of Iowa)
Najib El-Sayed (University of Maryland)
Jessica Kissinger (University of Georgia)
Larry Simpson (University of California, Los Angeles)
Andy Tait (University of Glasgow)
Marta Teixeira (University of Sao Paulo, Brasil)
Stephen Beverley (Washington University, St. Louis)

Executive Summary.

The family **Trypanosomatidae** of the protistan order **Kinetoplastida** includes three major lineages of human pathogens – the *Leishmania* and the African and American trypanosomes – that each rank within the top 10 in terms of global impact. The combined impact measured by DALY of these three parasites approaches 5 million. There are currently five ‘completed’ genomes of kinetoplastids: one African trypanosome, three *Leishmania*, and one American trypanosome, the latter being effectively a preliminary draft sequence due to assembly problems caused by its hybrid nature and extensive repetitive sequences.

Perhaps unlike the situation found in other groups of eukaryotic pathogens, trypanosomatids present two unique challenges. First is the fact that within each of the three major lineages, a wide range of disease pathologies is found. Thus, we are probing multiple diseases. Second is the fact that the trypanosomatids are relatively ancient, with divergences amongst the major lineages ranging from 200-500 million years ago. Thus, there is a wide range of evolutionary and pathological ‘space’ yet to be explored by genomic methods, accompanied by great opportunities for new insights and discovery.

The goals of this sequencing effort include improvement of: 1) our understanding of the mechanisms by which these pathogens cause disease; 2) our ability to influence the severity of these diseases through chemo- or immuno- prophylaxis and treatment; and, 3) our understanding of the biology that led these organisms to evolve into such successful pathogens.

A world community of several hundred investigators was queried as to the value of sequencing additional isolates from these three main groups of pathogens, and for advice as to which isolates are candidates for sequencing. Many dozens directly responded, indicating a high degree of interest and support. A consensus developed suggesting that isolates within *each* of these three major divisions should be sequenced with selections based on three criteria: 1) initial coverage of each of the major subgroups with a group; 2) coverage of closely related strains/isolates with different pathogenesis, and 3) coverage of strategically selected outgroups. Additional criteria include a focus on strains/isolates with well characterized phenotypes and/or that are being used by many laboratories in ongoing research.

The consensus of our respondents led to identification of 18 first priority isolates/strains for which genome sequence data would provide important new insights. A strategic six (6) species /

strains are recommended for completion. Additional recommendations were for limited cDNA end-sequencing and small RNA analysis from *L. braziliensis,* where RNAi pathways have recently been reported. As the genome size of trypanosomatids averages ~35 Mb, the target genomes total about 600 Mb. Second and third priority groups are also listed. The following tables summarize the proposal. The rationale for these selections is outlined in the text. Included in this list are projects planned and/or underway at the Wellcome Trust Sanger Institute (see letter appended from Matt Berriman), which will not require NHGRI resources.

Priority 1 *(proposed NHGRI target: 3 African, 6 American, 7 Leishmania, 2 outgroups)*

	Strain	Priority	gDNA	cDNA	Finish?
African Tryps					
T. b. brucei	Lister 427	1	10X	+	-
T. b. rhodesiense	Uganda	1	10X	-	-
T. evansi	American strain	1	10X	+	+
American Tryps					
T. cruzi TCI	Silvio X10	1	10X	+	+
T. cruzi TCIIa	Can III	1	10X	+	+
T. cruzi TCIIb	Esmeraldo	1	10X (2.5 X done TIGR)	-	-
T. cruzi TCIIc	3869	1	10X	-	-
T. cruzi TCIId	NRcl3	1	10X	-	-
T. cruzi TCIIe	Tula cl2	1	10X	-	-
Leishmania					
L. major	Friedlin	1	*Done*	+	*Done*
	Seidman	1	10X	-	-
L. tropica	TBD	1	10X	-	-
L. donovani	Strain selection in process by WTSI	1	10X planned at WTSI	-	-
L. mexicana	MHOM/GT/2001/U1103	1	5X started at WTSI	-	+
L. braziliensis (mucocutaneous)	TBD – 2 strains cutaneous, 2 strains mucocutaneous	1	10X	-	-
L. braziliensis	M2904	1	*Done*	+ & siRNAs	*Done*
L. panamensis	TBD	1	10X	-	-
Close Outgroups					
T. rangeli (for *T. cruzi*)	AM80/GpB	1	10X	+	+
Crithidia fasciculata (for *Leishmania*)	Cf-C1	1	10X; pilot underway at WU	+	+

Note: for gDNA, the X-fold coverage target is a target based upon paired-end sequencing of whole genome shotgun libraries. As new technologies are available, the target 'coverage' may be adjusted appropriately. For cDNA, 5' and 3' end sequencing of only 20,000 clones from a normalized cDNA library is planned due to the lack of introns in this taxonomic order (total 40,000 sequences reads from each of 8 species).

Priority 2 *(Proposed NHGRI targets: 2 American, 2 Leishmania)*

	Strain	Priority	gDNA
African Tryps			
T. b. gambiense	DAL 972 Type I	2	8X done at WTSI
T. b. gambiense	386 Type II	2	8X done at WTSI
T. congolense	IL3000	2	5X done at WTSI
T. vivax	Y486	2	5X done at WTSI
American Tryps			
T. cruzi TCI	Jose	2	10X
T. cruzi TCIIa	Tc11	2	10X
Leishmania			
L. aethiopica	TBD	2	10X
L. amazonensis	Marina/LTB0016	2	10X
L. tarentolae	Parrot-II	2	Pilot underway in Quebec

Priority 3 *(Proposed NHGRI targets: 2 American, 1 Leishmania)*

	Strain	Priority	gDNA
American Tryps			
T. cruzi TCIIb	Tc872	3	10X
T. cruzi TCIId	Tc656	3	10X
Leishmania			
Endotrypanum	TBD	3	10X
Outgroups			
C. deanei	library construction initiated and sequencing planned by WTSI	3	10X
Bodo saltans	library construction initiated and sequencing planned by WTSI	3	10X

The community endorses the NHGRI existing and standard data release policy for these projects as outlined at http://www.genome.gov/25521732

I. Introduction

A. Global impact of the trypanosomatid parasites.

The family **Trypanosomatidae** of the protistan order **Kinetoplastida** include three groups of human pathogens – the *Leishmania* and the African and American trypanosomes – that each rank within the top 10 in terms of global impact. The combined DALY of these three parasites approaches 5 million (WHO, 2002). The *Leishmania* are deadly pathogens that threaten millions of people in at least 88 countries, causing on the order of 1-2 million new cases annually. The African trypanosomes cause devastating public health challenges in Central Africa and livestock trypanosomiasis precluded development of hundreds of thousands of square miles of land that could be used for agriculture. *T. cruzi*, the American trypanosome that infects human populations, exerts the greatest impact in the Western hemisphere, where estimates of up to 20 million Latin Americans have different forms of Chagas Disease. Chagas disease is now emerging as a very serious problem in blood and tissue banks in the U.S. and Canada due to recent trends in immigration. Despite the overwhelming impact of these parasites, relatively little is understood about how they cause disease, and treatment typically employs drugs that show high toxicity and marginal efficacy.

The trypanosomatid parasites display a strikingly broad range of relationships with their hosts, and include non-pathogenic organisms (e.g., *T. rangeli*), extracellular parasites (e.g., *T. brucei*), intra-parasitophorous vacuole (e.g., *Leishmania* and *Endotrypanum*), and cytoplasmic (e.g., *T. cruzi*, *T. dionisii*) species. They infect a broad range of hosts, including all vertebrates groups (amphibians, reptiles, birds, fish, and mammals), many invertebrates (*Crithidia*, *Leptomonas*), and plants (*Phytomonas*). African trypanosomes are transmitted by tsetse flies, South American trypanosomes by reduvid bugs, and *Leishmania* by phlebotomine sand flies. Moreover, within these broad insect groups, different parasite species are transmitted by different insect species, a relationship which can often show considerable species selectivity. Notably, several of the kinetoplastid groups closest to the trypanosomatids (e.g., *Bodo*, *Cryptobia*, *Trypanoplasma*) are free-living. These flagellates also present extremely unique and unusual molecular systems, including the novel kinetoplast and glycosome organelles, antigenic variation of surface glycoproteins, polycistronic transcription, trans-splicing of mRNAs, RNA editing, etc., that participate at some level in their ability to cause pathogenesis and disease.

Although the genetic bases of the phenotypes or these organisms are only poorly understood, they certainly represent potential targets for chemo- or immuno- prophylaxis or therapy. Comprehensive genomic analysis of selected members of this clan of important parasites will provide an excellent step toward a better understanding of their biology and pathogenesis. Herein, we focus our attention on the highly pathogenic isolates of the American and African trypanosomes and the *Leishmania*.

B. Outstanding questions in the pathogenesis of kinetoplastid parasites.

1. *What is/are the genotypes associated with the ability of different strains or isolates to cause widely varied clinical manifestations? Are certain metabolic, regulatory, or genetic networks required for or associated with disease?* Each of the three main groups/species of trypanosomatids pathogenic for man presents diverse clinical outcomes. Different species of *Leishmania* cause disease ranging from self-resolving cutaneous symptoms to lethal systemic manifestations. Chagas Disease also presents with a wide variety of clinical outcomes, including chronic chagasic cardiomyopathy, the ‘mega’ syndromes, or even totally asymptomatic carriers, and many patients do not manifest disease until years after the infection. The African trypanosomes range from non-infectious for man, to causing long-term chronic infections, relapsing disease or rapid and lethal infections. The genetic bases of the diversity of clinical outcomes in these parasites are largely unknown. Genomic dissection of a select group of these parasites will provide a window into the genetic basis of these pathogenic characteristics.

Which isolates will provide this information?

a. The salivarian / African trypanosomes. *T. b. brucei* vs *T. b. gambiense* vs *T. b. rhodesiense.* Each of these *T. brucei* subspecies exhibits specific pathogenic / clinical manifestations. *T. b. brucei* is generally non-infectious for people; whereas *T. b. gambiense* and *T. b. rhodesiense* isolates from western and eastern Africa exhibit strikingly variable clinical manifestations and pathologies in infected humans. *T. evansi*, which is generally non-infectious for humans, but is much more broadly encountered (it is the only 'African' trypanosome found outside of Africa) and is pathogenic for many mammalian hosts. Comparison of it's genome with those of the *T. brucei* species and *T. congolense* and *T. vivax*, being sequenced at the Sanger Center will provide a basis for identifying the genetic roots of these differences.

b. The sterchorian/ American trypanosomes. It has been widely shown that the two major lineages of *T. cruzi*; i.e., TCI and TCII, exhibit significant differences in pathogenic potential. TCI is generally less pathogenic for humans, has a lower acute infectious profile and progression, a more extensive chronic profile, and invades and causes pathology in different organs. Comparison of at least one TCI isolate (e.g., Silvio X10) with the other TCII isolates will provide an opportunity to discover the genetic basis of these phenomena.

c. The *Leishmania.* As Leishmania sp. are thought to have originated approximately within the last 100 MYR, comparisons amongst all species will be useful in addressing these questions. Comparisons amongst closely related species provide the greatest signal; these include in the Old World, *L. major* Friedlin vs. Seidman, *L. tropica* and *L. aethiopica* (healing vs. various degrees of chronic cutaneous disease), in the New World, *L. mexicana* v. *L. amazonensis,* and *L. braziliensis* strains and *L. panamensis* (mucucotaneous vs. cutaneous leishmaniasis).

2. *Can we identify genetic determinants or complements that are associated with the ability to invade and replicate in different intracellular compartments?* As mentioned above, the site of replication of each of these groups of trypanosomatids differs. The African trypanosomes (and a subset of American trypanosomes) replicate exclusively extracellularly. The *Leishmania* replicate in parasitophorous vacuoles in the cytoplasm of infected mammalian macrophages. *T. cruzi* replicates freely in the cytoplasm of the infected cell, and can infect many different mammalian cell types. Moreover, evidence is clear that different isolates of the American trypanosomes, as well as different species of *Leishmania,* infect susceptible mammalian cells via multiple and often mutually exclusive pathways. Genomic comparisons among several isolates exhibiting each of these phenotypes would provide clues to the genetic basis of the ability to occupy a given cellular compartment.

Which isolates will provide this information?

As described in the paragraph above, all African trypanosomes are extracellular, and the American trypanosomes and *Leishmania* replicate in different cellular environments. Thus, the comparisons that will provide this information are inter-generic. Patterns of genetic and metabolic pathways that are associated with the abilities to exist in these different cellular compartments will be identified by comparing the pools of genes in the African, American and *Leishmania* parasites. Comparisons involving *L. tarentolae* (lizard

Leishmania) and *Endotrypanum* (residing at least in part within red blood cells) probe the role of temperature and/or cellular location.

3. *Are there common genetic characteristics associated with the very different routes of transmission displayed by these parasites?* The kinetoplastid parasites exhibit extraordinarily diverse vectors and modes of transmission. Whereas almost all species are transmitted by insects, the African trypanosomes are considered 'salivarian' since they are transmitted by inoculation of infective forms from the salivary gland of the infected tsetse fly. *Leishmania* are inoculated into the host during a bite of an infected sand fly. In contrast, the American trypanosomes are considered 'sterchorarian' or transmitted via a contaminative route after the bite of a triatomine bug. However, the American trypanosome *T. rangeli* is apparently transmitted via either the inoculative or the contaminative routes. *T. evansi* and *T. equiperdum* are considered 'African' trypanosomes, but they are only transmitted mechanically by biting flies or venereal contact, resp. Functional genetic reconstruction and comparison of selected species and strains of these parasites would clarify the genetic basis of these characteristics.

   Which isolates will provide this information?

   a. African Trypanosomes. All of the *T. brucei* subspecies (*brucei*, *rhodesiense*, and *gambiense*) are transmitted from the salivary gland of the tsetse fly. *T. evansi* is transmitted mechanically. *T. rangeli*, which can provide the function of an outgroup for the African trypanosomes (although it is closer to the American trypanosomes), is transmitted *either* from the salivary gland of infected insect, or mechanically. Thus, comparison of these isolates should provide information on the genetic basis for these routes of transmission.

   b. American Trypanosomes. All of the TCI and TCII lineages and sublineages are transmitted by deposition of feces from an infected insect into a bite wound. *T. rangeli*, as previously described, is transmitted from the salivary gland of a biting insect, or by deposition of feces into a bite wound. The existence of these two routes of transmission in *T. rangeli* makes it an ideal comparison with the *T. cruzi* isolates, as well as with the African trypanosomes.

   c. The *Leishmania*. *Leishmania* species are transmitted by the bite of phlebotomine sand flies, either the genus *Phlebotomus* (Old World) or *Lutzomyia* (New World). These genera and species therein differ greatly in their interactions with different *Leishmania* species, for example in midgut lectins important to binding and immunomodulatory salivary gland secretions. *Leishmania* of the subgenus *Viannia* (*L. braziliensis*, *L. panamensis*) reside within the hindgut rather than the midgut. Comparisons amongst *Leishmania* isolates along these lines would reveal candidate genes implicated in these processes.

4. *What is the genetic basis of the ability of these organisms to parasitize and inhabit diverse ecological habitats? How did parasitism evolve in this most successful group of parasites?* The kinetoplastids represent one of the most successful protozoan groups with free-living species inhabiting a range of ecological niches, and parasitic species infecting vertebrates and invertebrates of several classes, in which they display a broad variety of life cycle types. The species of *Bodo* are free living; the African trypanosomes (and several other trypanosomes; e.g., *T. rangeli*) are exclusively extracellular, but the pathogenic American trypanosome (*T.*

cruzi) and the *Leishmania* spp. grow exclusively intracellularly in their vertebrate hosts. Genomic study of these organisms could identify common gene complements or genetic systems associated with these diverse life styles.

Which isolates will provide this information?

> A cross genus comparison the genes and systems common to isolates from the African and American trypanosomes, and the *Leishmania*, which have diverged significantly from their free-living predecessors like the *Bodonids*, will identify those genes and systems (presence, or lack thereof) associated with a parasitic life style. Whereas this would best be performed using a free-living kinetoplastid (e.g., *Bodo saltans*), significant insight will be gained by study of and comparison of the genes and systems retained in or lost from the three pathogenic groups (e.g., the African and American Trypanosomes, and the *Leishmania*), as well as the outgroups *T. rangeli*, *T. evansi*, and *Crithidia fasciculate*). Other comparisons with *Bodonids*, *Cryptobids*, and other members of the Phylum Euglenozoa would also be helpful, but may be beyond the scope of this project.

5. *How did the ancestral free-living parasites evolve into the current-day free-living bodonids and cryptobids, and simultaneously evolve into extremely successful parasites? How is pathogenicity selected, or what are the selective advantages provided by a pathogenic/parasitic life style?* Though closely related to question 4 above, this query relates to the fact that all trypanosomatids are parasitic. However, it is likely that they evolved from free-living flagellate and that the parasitic life style originated more than once. A broad comparison of the genomes of several of organisms from each of the main groups, including the free-living bodonids, could answer the question of why this group of protozoa evolved into one of the most successful groups of parasites known. These questions have broad impacts on our understanding of the natural history and epidemiology of parasitic diseases.

   Which isolates will provide this information?

   > The best comparisons for this purpose would be provided by and examination of the genomes the free-living outgroups to the pathogenic groups. Thus, parasitism obviously has developed in several different pathways from the free-living ancestral trypanosomatids to the three major pathogenic subgroups. A comparison therefore, of the genomes of the free-living bodonids and cryptobids, (e.g., *Bodo saltans*) with the genomes of the pathogenic species, might well provide clues as to how these parasitic life styles evolved. Since the free-living species have been reduced in priority due to our focus on pathogenic isolates, information could well be provided by comparison of the genomes of the selected outgroups: e.g., *T. rangeli*, *C. fasciculata,* and in some ways *T. evansi*, with the genomes of the pathogenic species: e.g., *T. cruzi*, the African trypanosomes, and the *Leishmania*, respectively.

6. *What are the genetic bases of the phenotypic (cell cycle, host range, vector selection, pathogenic and clinical manifestations, etc.) characteristics of the major groups of pathogenic trypanosomatids?* Each of the isolates selected for sequencing is a member of a subgroup with specific characteristics. Many of these are shared by other groups, but many are not. The genetic basis of these differences would explain why these trypanosomes are responsible for different disease states. A comparison of their genomes is probably the only way to obtain this information.

Which isolates will provide this information?

a. African isolates. *T. b. brucei* as mentioned above, generally infects only cattle, whereas *T. b. rhodesiense* and *T. b. gambiense* infect primarily humans. *T. congolense* and *T. vivax* largely do not infect humans, but rather infect ruminants or pigs. *T. evansi* and its close relative *T. equiperdum*, tend to infect horses, donkeys, mules and other large mammals, but not humans. Moreover, the latter two species are spread by mechanical transmission, in contrast to the other African trypanosomes that are spread by salivarian routes. Finally, there are significant differences, as pointed out above, in the pathogenic profiles of *T. b. brucei*, *T. b. gambiense*, and *T. b. rhodesiense* (not to mentione *T. evansi*). Thus, a comparison of these closely related organisms will provide a guide to their major phenotypic differences.

b. American isolates. Isolates of the two lineages of *T. cruzi* are quite divergent in many respects (see below), and have been argued to represent members of different species. Although superficially similar, their preferred hosts and vectors, method of invasion, affects on the invaded cells, levels of parasitemia, mechanisms of pathogenesis, and clinical outcomes are quite different. Similarly, within the most commonly studied group of *T. cruzi*, lineage TCII, there are now five clearly separated subgroups (a-e). Each of these subgroups exhibits distinct characteristics, in terms of life cycle and pathological features. Whereas it is quite well established that the differences among TCI and TCII isolates are genetically programmed, it is not yet established which genes or gene networks confer these different phenotypes. Thus, obtaining a good draft sequence, with the subsequent gene annotation and metabolic and other network reconstructions of these isolates (TCI Silvio X10, TCIIa Can III, TCIIb Esmeraldo, TCIIc 3869, TCIId NRcl3, and Tula cl2) will provide a basis for a comparison that will identify the genetic roots of these differences.

c. The *Leishmania*. Genetic differences associated with the transition from insect only to insect + mammalian parasitism will be visualized by comparisons of *Crithidia fasciculata* and *C. deanii* against *Leishmania* species. Differences associated with mucocutaneous v. cutaneous leishmaniasis will be established by comparisons amongst *L. braziliensis* isolates from well characterized patient isolates, and also against *L. panamensis*. Differences associated with disease pathology in old world cutaneous leishmanisis will be obtained by comparisons of L. major Friedlin (healing), *L. major* Seidman (nonhealing, chronic), *L. tropica* and *L. aethiopica*. Differences in vector/transmission will be assessed by comparisons described earlier.

7. *What are the signals used to control gene expression?* Trypanosomatids are unique in employing a novel polycistronic transcriptional mechanism to express their genes. At present the signals which key aspects of RNA processing such as trans-splicing, or RNA stability and/or translation, are not well understood. Comparative analysis amongst appropriately divergent lineages is now an established approach to identifying these critical signals. Although the three major lineages of trypanosomatids have proven too distant for this purpose, within each lineage this powerful methodology is readily applicable.

Which isolates will provide this information?

One of the main weaknesses with the existing genome sequences of the trypanosomatids is that gene identification is very weak, and identifications of accurate start and stop signals, splicing signals, and other regulatory information is nearly non-existent. Current EST banks are limited and poorly representative. Because transcription in this organisms is polycistronic, it is nearly impossible to use the existing information to find relevant gene control information, much less identifying the likely start and stop codons, potential *trans*-splice acceptor sites, or even upstream or downstream non-translated regions. Thus, we have selected a small number of isolates (2 from each of the major groups, in addition to the two outgroups), for cDNA sequencing. A powerful advantage arises from the lack of *cis*-splicing; transcripts can be mapped solely by 5' and 3' end sequencing, an approach that should be amenable for adaptation to high throughput 454 or other methodology. Availability of this information for the few isolates described will very likely permit identification of signals in related isolates.

a. African Trypanosomes. For the African strains, we have selected the type strain *T. b. brucei* Lister 427, and *T. evansi*, which is divergent from the other African trypanosomes for this analysis.

b. American Trypanosomes. To date, the existing EST data available for *T. cruzi* isolates is extremely limited. We recommend sequencing cDNA from the type strain of TCI (Silvio X10), and the type strain of TCII subgroup a (Can III). The combination of data from these two isolates should permit characterization of the genomes of most of the other strains to be sequenced.

c. The *Leishmania*. As for the American trypanosomes, existing EST data is limited the *Leishmania*. We have suggested that two important isolates, the Friedlin strain and M2904 of *L. braziliensis* whose genomes have already been sequenced, be analyzed through transcript mapping by EST end sequencing. This, again, will provide information that will probably be relevant to all of the *Leishmania* sequenced.

d. Outgroups. We suggest EST end-sequencing sequencing of each of the two high priority outgroups, *T. rangeli* and *C. fasciculata.* This data will be required to compare these genomes to the others of the pathogenic species. That is, a good annotation requires the EST libraries, without which, a valid comparison with the other species of trypanosomatid will be a challenge.

C. Evaluation criteria for strain/species for study.

Perhaps unlike the situation found in other groups of eukaryotic pathogens, within the trypanosomatids we are confronted with two unique challenges. First, as described above, within each of the three major lineages, a great range of disease pathologies are found. It is widely believed that genetic differences amongst the parasites strains are the major cause, although host factors also clearly play a role. For example, within the genus *Leishmania* there are numerous disease pathologies ranging from cutaneous to mucocutaneous to fatal visceral leishmaniasis, each of which is associated with a subset of parasite species and/or strains. Similarly different forms of *T. cruzi* and *T. brucei* have been associated with differing pathologies in humans, domestic or wild animals. Second is the fact that the trypanosomatids are relatively ancient, with divergences amongst the major lineages ranging from 200-500 million years ago. Thus there is a wide range of evolutionary and pathological 'space' yet to be explored by genomic methods.

Many different criteria have been used to select organisms for genome sequence analysis. Our goal is to use the genomic data generated to better our understanding of the impact of these parasites on humans through there ability to cause disease. Thus, we have selected strategically placed species, isolates or strains, for which the genomic sequences and the information thereof derived, will likely provide the maximum impact on:

1. our understanding of the mechanisms by which these pathogens cause disease;
2. our ability to influence the severity of these diseases through chemo- or immuno-prophylaxis and treatment; and,
3. our understanding of the biology that led these organisms to evolve into such successful pathogens.

To accomplish this, we propose to select species according to the following principles:

1. coverage of **major subgroups** within a group of parasites;
2. coverage of **closely related strains**/isolates with clearly different pathogenesis; and,
3. coverage of **strategically selected 'outgroups'**.

Major subgroups. First, we propose to provide a comprehensive coverage of major groups or lineages associated with these pathogens. Thus, these parasites have been extensively characterized and each of the major pathogenic groups (e.g., the *Leishmania* and the African and American trypanosomes) has been subdivided into taxonomic/phylogenetic groups. In many cases, disease related phenotypes follow these groups. For example, it is generally accepted that *T. cruzi* in Group I are much more prevalent in the sylvatic cycle than in the domestic cycle in all of Latin America but that this lineage predominates from Mexico to Northern Brazil (Amazonia) as the agent of human disease. Due to the absence of the mega-syndromes, Chagas' disease caused by *T. cruzi* group I isolates is considered less severe than infections with *T. cruzi* group II isolates.

Closely related isolates with differing pathogenic phenotypes. Second, we propose to analyze isolates that are clearly closely related but that cause significantly different disease or display other disease relevant phenotypes that have a likely genetic basis. Each of the parasite groups studied exhibits examples of isolates that are clearly closely related genetically (e.g., based on phylogenetic analysis of the rRNA or other gene sequences), but that display striking differences in pathogenicity. Previous pathogen sequencing efforts have shown that comparisons of closely related strains and/or species are able to link specific genes or alleles with the disease process. For many of these the presence of sexual exchange and genetic linkage allows one to associate specific alleles/haplotypes with phenotypes. This is a relevant factor as trypanosomatid parasites show wide variety in the frequency of sexual exchange, from *T. brucei* which has a well characterized meiotic process to *Leishmania* where sexual exchange is at best relatively rare and potentially nonexistent. Nonetheless, absence of sexual exchange does not preclude harvesting important information by genome comparisons, one example being comparisons of dN/dS ratios as evidence for selection. In short, comparison of closely related isolates with different phenotypes has the potential to identify the genetic basis of these differences and lead to a much better understanding of the pathogenic determinants responsible for disease.

Strategically selected 'outgroups'. Third, we propose to study selectively informative outgroups that are closely related to, but different from, the highly relevant human pathogens. Again, each of the major pathogenic groups exhibits closely related species that, though parasitic, are apparently non-pathogenic for humans; e.g., *T. cruzi* and non-pathogenic *T. rangeli*. These may represent differences in the human disease pathology (cutaneous, mucocutaneous or visceral leishmaniasis, or non-pathogenic *T. rangeli* and *T. cruzi*), in human versus mammalian infection (*T. b. brucei*, *T. b. rhodesiense* or *T. b. gambiense*, or *Sauroleishmania (L. tarentolae*) and mammalian *Leishmania*), or vertebrate versus insect parasitism (*Leishmania* vs. *Crithidia*). Study of these organisms has the potential to provide insight into the genetic determinants of human-host infectivity of these parasites.

We also considered the study of a free-living flagellate that is most closely related to the pathogenic trypanosomatids as the most closely related non-pathogenic outgroup (*Bodo* versus *T. brucei, T. cruzi* and *Leishmania*). This comparison would have the potential to identify genes associated with the transition to parasitism. Its value is compromised to some extent by the fact that this is an ancient divergence, and many differences detected may reflect evolutionary noise. While the collective opinion was that would not compromise its utility, it prompted us to deprioritize this relative to other genomes.

<u>'Laboratory vs. wild'</u>. Considerable thought was given to the nature of the specific strains within each species or type to be selected for sequence analysis. Two criteria were debated: how close to the 'field' should the strains be, and whether strains should have been demonstrably shown to undergo the complete life cycle experimentally. **As discussed below, the collective opinion was that the focus should be on well-characterized isolates that had undergone very limited culture *in vitro* prior to DNA acquisition.** This of course does not preclude study in some instances of strains derived decades ago, but stored at -150º C. All strains to be studied here are low-passage isolates.

Ideally, we prefer to study strains exactly in their 'wild' state; e.g., without changes associated with adaptation to culture. While abundant data attest to the ability of trypanosomatid genomes to change during extended culture *in vitro* in diverse ways, as of yet there are no data supporting the notion that significant changes occur between 'wild' parasites and ones that have only recently been cultivated. Many (not all) trypanosomatids are readily cultured when taken from their animal or insect hosts, which could support the notion that significant genomic change is unnecessary. However, in reality, the situation is best characterized as an 'absence of evidence' than 'evidence of absence'. Collectively we judged the risk associated with working with recently cultivated, low passage isolates to be low. The availability of extensive genome sequence data arising from these studies will facilitate explorations of this question in the future.

A second question was whether all strains studied should be experimentally transmissible through the entire life cycle, including animal models showing pathology relevant to that seen in humans. While obviously a useful and a positive criterion for inclusion, the collective judgment was that this should not be a litmus test for inclusion. For African trypanosomes it is generally not possible to carry a single isolate around the entire life cycle. While possible for some *Leishmania* and American trypanosomes, it is not true for all. Moreover, for all species the relevancy of the animal models is often questionable; e.g., there are no accepted models for mucutaneous or certain other forms of leishmaniasis and the models for visceral leishmaniasis are far from ideal.

II. Strain/isolate selection.

Current State of kinetoplastid sequencing. The current state of kinetoplastid sequencing projects is included in the candidate strain lists below. These isolates can be classified according to: African or American trypanosomes, or *Leishmania*. Four African trypanosomatids are in various stages of completion. Three species of *Leishmania* have been completed, and two others have been initiated (pilot or otherwise). No American trypanosome has yet been fully-sequenced. *T. cruzi* CL Brener has been extensively sequenced but its assembly and use has been a major challenge due to its hybrid nature, and only ~2.5 coverage of *T. cruzi* Esmeraldo is available. A partial sequence of a bat trypanosome, a possible representative of a progenitor of *T. cruzi*-like isolates, has been obtained. Pilot projects for one trypanosomatid from insects (*Crithidia fasciculata*) and one from plants (*Phytomonas* sp Jma) are in progress. In short, the state of genome sequencing of the kinetoplastida is spotty and largely incomplete.

Several projects have been initiated, planned and/or contemplated by the Wellcome Trust Sanger Institute (WTSI) for various trypanosomatids, discussed individually in the following sections. The group felt strongly that priorities should be discussed separately from the question of who would take on specific genomes. Having done so, it was also agreed that duplication and/or overlap in effort should be avoided. We have included with the white paper a supporting letter outlining the current WTSI projects and plans relevant to trypanosomatids; this information is also included in various tables throughout this document. It will be evident that there is no overlap, and great potential for synergy.

A. African Trypanosomes. The African trypanosomes traditionally include members of the *T. brucei* clade. This clade consists of all African trypanosomes of the Salivaria Section, with species classified in the subgenera *Trypanozoon* (*T. brucei ssp, T. evansi, T. equiperdum*), Duttonella (*T. vivax*), Nannomonas (*T. congolense*) and Pycnomonas (*T. suis*). African trypanosomiasis is a major constraint to human and livestock development throughout the tropical regions of Africa where tsetse flies are prevalent, and in America and Asia where some species have adapted to mechanical transmission. Most phylogenetic analyses suggested an early divergence of African trypanosomes (at least 100 mybp (million years before present) when Africa became isolated from other continents), and an evolutionary history confined to Africa and associated with tsetse. The tsetse-transmitted trypanosomes of African mammals group together in an exclusive clade (including *T. evansi* and *T. equiperdum*, which are probably recently derived from *T. brucei*). Analysis of 18S rRNA sequences places *T. brucei* in a clade comprising exclusively mammalian trypanosomes of African origin, suggesting an evolutionary history confined to Africa. These species are, in general, pathogenic for their mammalian hosts. Human infective trypanosomes comprise two sub-species (*T. b. gambiense* and *T. b. rhodesiense*) with different geographical distributions and courses of infection. Antigenic variation allows African trypanosomes to develop chronic infections in mammalian hosts. Most species are transmitted by inoculation of tsetse saliva. Exceptions are *T. vivax*, *T. evansi* and *T. equiperdum,* which are mechanically transmitted and thus are the only African trypanosomes

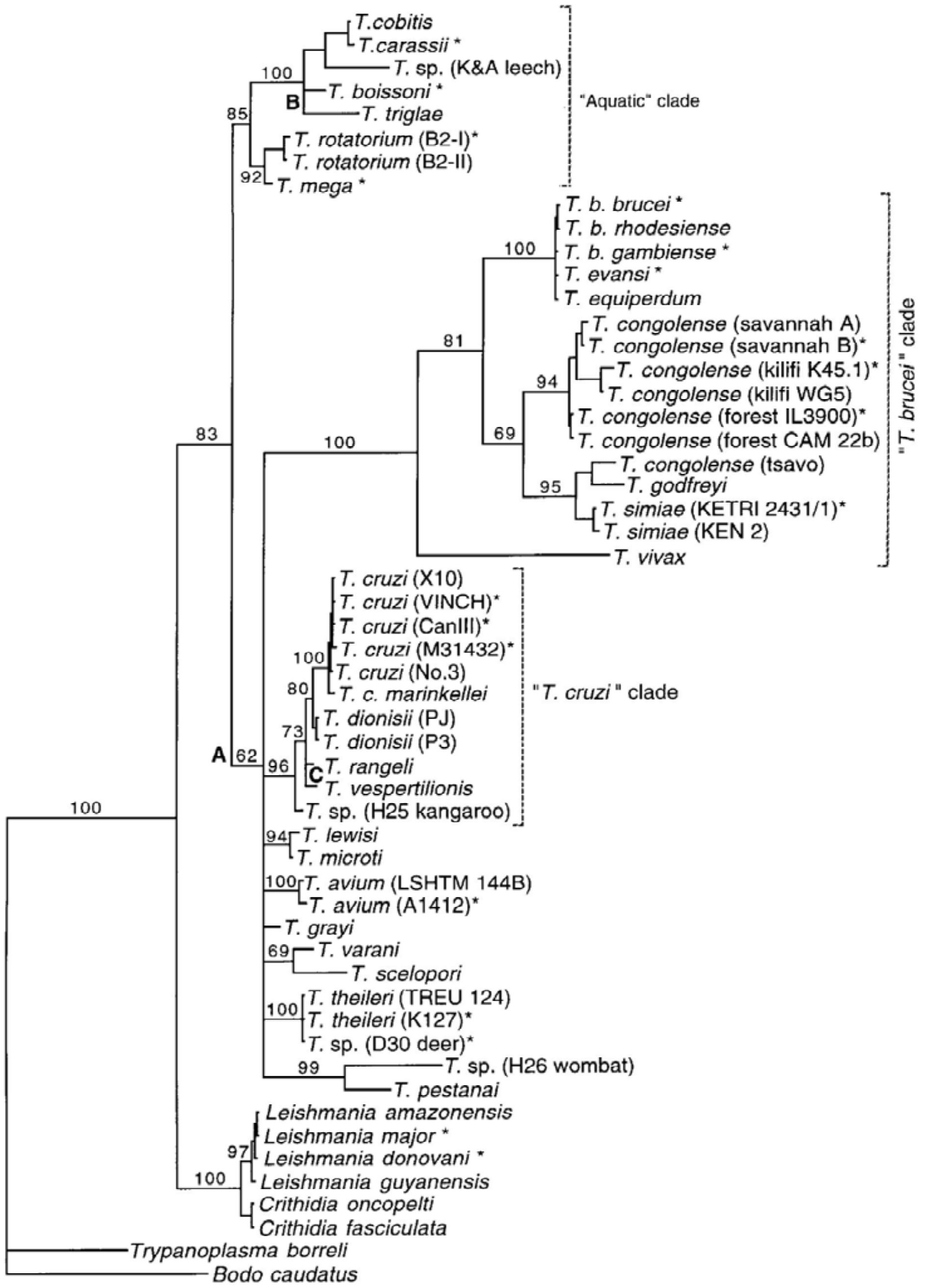


Fig. 1. Phylogram constructed by bootstrapped (100 replicates) maximum parsimony analysis of 55 kinetoplastid 18S ssu rRNA sequences. From Stevens et al., 1999.

outside Africa. *T. vivax* can also be transmitted through inoculation by tsetse flies, or mechanically by other biting flies. *T. evansi and T. equiperdum* do not develop in invertebrates because they lack functional mitochondria.

A. *T. brucei* ssp. *T. b. brucei* infects various species of several artiodactyl families as well as other domestic and wild mammals causing livestock disease. *T. b. rhodesiensi* and *T. b. gambiense* circulate between domestic livestock and man and are responsible for human African trypanosomiasis (sleeping sickness). Historically, epidemic sleeping sickness has caused massive loss of life, and related animal diseases have had a crucial impact on development in sub-Saharan Africa. *T. b. rhodesiense* and *T. b. gambiense* are morphologically indistinguishable from *T. b. brucei*, which does not infect humans because it is lysed by a lytic factor associated with High Density Lipoproteins (HDLs) in normal human serum. As shown in the figure above, the *T. b. brucei, T. b. gambiense, T. b. rhodesiense, and T. evansi* are quite closely related. However, these species exhibit striking differences in host range, infectivity, and disease progression (e.g., *T. b. rhodesiense* generally causes acute infection in humans, in distinct contrast to *T. b. gambiense*). Thus, the analysis proposed is designed to identify the genetic basis of the clear phenotypic differences among these strains.

Species	**Strain**	**Host**	**Geog.**	**State of Sequence**	**Genet**	**Priority**	**Avail**
T. b. brucei	TREU 927	Tsetse	Kenya	Mostly complete with a few gaps	+	low	yes
T. b. brucei	Lister 427	Ovine	Uganda	Not done	+?	1	yes
T. b. gambiense	DAL 972 Type I	Human	Africa	8 fold at Sanger	+?	2*	yes
T. b. gambiense	386 Type II	Human	Africa	Planned at Sanger	+	2*	yes
T. b. rhodesiense	Uganda isolate	Human	Uganda	Not done	+?	1	yes
T. congolense	IL3000	Bovine	Kenya	5 fold at Sanger	unk	2*	yes
T. vivax	Y486	Bovine	Nigeria	5 fold at Sanger	unk	2*	yes
T. evansi	American strain	Broad	Broad	Not done	-?	1	yes

***** While identified as a high priority target by the writing group, these strains are being sequenced at the Wellcome Trust Sanger Institute (WTSI), and thus were not targeted for inclusion by the NHGRI effort. . +? suggests that genetic exchange has been demonstrated in closely related strains. -? it is generally thought that genetic exchange is unlikely in *T. evansi*.

***T. b. brucei* TREU 927**. This isolate of *T. b. brucei* has been sequenced and published. The Sanger Institute has on ongoing project to close the remaining gaps in the large chromosomes. No additional efforts are recommended at this time.

***T. b. brucei* Lister 427.** Almost every question about the biology of *Trypanosoma brucei* is being addressed in Lister 427 clones, in almost every trypanosome lab worldwide. There are some important biological differences between this and TREU 927, the ‘genome strain’, which might be explained via genomic sequencing. Sequencing would also contribute greatly to understanding the evolution of gene families, especially in the context of antigenic variation, which is the greatest contributor to virulence of the African trypanosomes. Various clones of this strain differ in regard to tsetse transmissibility: efficiently transmissible and non-transmissible clones are available. A relatively recent clone, KMCBM, has retained good transmissibility. Although transmissibility variation is possibly due to phenotypic variation, it could equally be genetic, in which case sequencing of non-transmissible clones may identify genes that are needed to complete the life cycle. Reports of significant differences in gene organization (both copy number and point mutations in individual genes) between Lister 427 and TREU 927 suggest that a complete sequence could be informative about gene evolution and assistive to genetic manipulation that is still most readily performed in this strain. The isolate should be sequenced as completely as possible as a service to the community and as an ideal comparative to the existing sequencing of TREU 927. Whereas this strain was isolated many years ago, it has been extremely well characterized in terms of its biology, genetics, and pathogenicity, and it would be a disservice to the community not to include it in this priority listing.

***T. b. gambiense* DAL 972 (type I) and 386 (type II)**. *T. b. gambiense* has been subdivided into two 'types' (1 and 2) on the basis of isoenzyme and molecular markers. They exhibit less than 1% divergence from *T. b. brucei* at the nucleotide level in the SSU rRNA. An isolate of each of these types (from the Cote d'Ivoire) is being sequenced at the Sanger Center. The species have been shown to undergo sexual reproduction and thus should be sequenced.

***T*. *rhodesiense* (a Uganda strain)**. To date no genome project has been undertaken on *T. b. rhodesiense*. There are substantial, important differences in the biology of the three subspecies of *T. brucei* in host range, mechanisms of human-infectivity, tsetse species used as vectors, virulence and drug sensitivities. *T. b. rhodesiense* isolates exhibit less than 1% divergence from *T. b. brucei* at the nucleotide level in the SSU rRNA. A number of population studies of *T. b. rhodesiense* coupled with the clinical and epidemiological analysis has shown that: (1) each focus of disease appears to be genetically distinct with greater similarity to local *T. b. brucei* than to *T. b. rhodesiense* strains from a different focus; (2) in one focus which was studied in detail, the same clonal genotypes have persisted for >30 years; (3) the severity of disease differs between foci with a less virulent form of the disease being observed in Zambia and Malawi compared to the virulent disease observed in Uganda and Kenya; (4) the disease occurs in epidemics with sudden expansions of the number of cases followed by periods of low prevalence, (5) there are host genotypes that potential determine the severity of disease. While we understand some of the factors that lead to these differences, such as the SRA gene that is a determinant of human infectivity in *T. b. rhodesiense*, it would be understatement to say that this understanding is fragmented. Genomic sequence and comparison with sequences of other *T. brucei* species would address the genetic bases of these differences, clarify the organization of VSG genes in *T. b. rhodesiense*, and identify genes that are under positive selection among these subspecies.

T. evansi. *T. evansi* has a wider geographical distribution and mammalian host range than any other pathogenic trypanosomes, infecting the horse, camel, buffalo, cattle, pig, deer, dog, and several wild mammals. Though generally non-infectious for humans, a case of *T. evansi* infection in a human was recently reported in India (Am J Trop Med Hyg. 2005 73: 491-5; N Engl J Med. 2006 355: 2752-6). However, the serum of the infected patient was found to have no trypanolytic activity, and the finding was linked to the lack of APOL1, which was due to frameshift mutations in both APOL1 alleles. This species is also only transmitted mechanically rather than from the salivary gland of an infected tsetse. It causes a wasting disease in Africa, South America, and Asia. Despite these seemingly gross biological differences, the SSU rRNA sequence of *T. evansi* is less than 1% divergent from that of *T. b. brucei* and *T. b. rhodesiense* and it is thought to have arisen recently from *T. brucei* (see fig. 1). All molecular markers investigated demonstrate that *T. evansi* has the highest genetic homogeneity among all species of the subgenus *Trypanozoon*. No sequencing efforts are ongoing for this species and it was thus be prioritized highly in this program. The strain selected is from the Americas to increase the diversity (all the other isolates from this group are from Africa), and has been passaged only limited times prior to creating frozen stabilates in the TCC culture collection maintained by Marta Teixeira.

***T. congolense* IL3000.** *T. congolense* infects bovids, camels and pigs, but is not infective for humans. In addition to this species, the subgenus Nanomonas includes *T. simiae* (restricted to pigs and agent of an acute fulminating disease), which proved to be quite different from *T.*

congolense by biochemical and genetic analysis. However, these species form a group that is well distinguished from the *T. brucei* clade (see fig. 1). The SSU rRNA sequence varies by ~15% from *T. b. brucei*. Sequencing of *T. congolense* will address the genetic source of the host selectivity of these trypanosomes. The Sanger Institute is producing a 5X coverage and an EST library of *T. congolense.* Thus, completion of the genome sequencing of this isolate should be prioritized by NIHGRI. Eventually, as there are three divergent subtypes (Savanna, Forest and Killif), additional isolates of *T. congonlense* will need to be sequenced.

***T. vivax* Y486.** *T. vivax* is mostly parasitic of ruminants and is not infective for humans. Disease ranges in severity from chronic and asymptomatic to acute and fatal. Stocks from West, East and Central Africa differ in morphology, growth, pathogenicity, and molecular markers. *T. vivax* is one of the less well studied African trypanosomes, with controversial phylogeny. Its placement within the *T. brucei* clade makes sense as adaptations requiring complex developmental pathways and sophisticated genetic machinery as antigenic variation are unlikely to have evolved independently in different trypanosomes. However, several characteristics; e.g., GC content, secondary structure of SL RNA, chromosomal organization, etc., suggest a wide evolutionary distance from other members of the *T. brucei* clade and *T. congolense* (see fig. 1). The sequence of the SSU rRNA varies by nearly 20% from that of *T. b. brucei*. The Sanger Institute is completing a 5X coverage of *T. vivax* Y486 (from West Africa). Thus, completion of the genome to a high depth should be prioritized by NIHGRI.

B. South American trypanosomes. South American trypanosomes customarily include the species *T. cruzi*, *T. rangeli* – the only two species that infect humans in the Americas, and several related parasites (e.g., *T. dionisii*) isolated from bats and other mammals. *T. cruzi* (from humans and sylvatic mammals) clusters with trypanosomes specific to Old and New World bats, *T. rangeli* and a trypanosome species isolated from an Australian kangaroo. Interestingly, although *T. cruzi* and *T. rangeli* are found in South and Central America, where they infect a broad range of sylvatic and domestic mammalian hosts (WHO, 1997), two other species (*T. dionisii, T. vespertilionis*) are specific to bats and are found throughout the Americas and Europe. Although a long list of related species has been identified, there is no evidence that these are infective for humans. *T. dionisi,* one of the best outgroups has been sequenced to a 10X genomic coverage with ~40,000 EST clones (Buck et al., unpublished). Therefore, for the purposes of this proposal, we will focus only on the characterized isolates of *T. cruzi* and *T. rangeli*. We will apply the three primary criteria outlined above, tempered by the availability of some well-characterized type strains.

B. Clade *T. cruzi* (see below under 'Outgroups' for Clade *T. rangeli*). *T. cruzi* infects all mammalian orders, including man, and is transmitted by triatomine bugs throughout Latin America. The parasite emerged ~150 mybp, originally infecting primitive mammals, but only recently -- 30,000-15,000 ybp -- had first contact with humans. Dozens of *T. cruzi* isolates are available.

Molecular markers define two major phylogenetic lineages, *T. cruzi* 1 (TCI) and *T. cruzi* 2 (TCII). Interestingly, analysis of the SSU rRNA genes of these organisms shows that the genetic distance between TCI and TCII is approximately the same as the genetic distance between the *Leishmania* and the *Crithidia*. TCI is considered to be comparatively homogeneous whereas TCII is made up of five subgroups (a-e) [see fig. 2]. These subgroups show clear and consistent divergence using various phylogenetic metrics (e.g., SSU rRNA

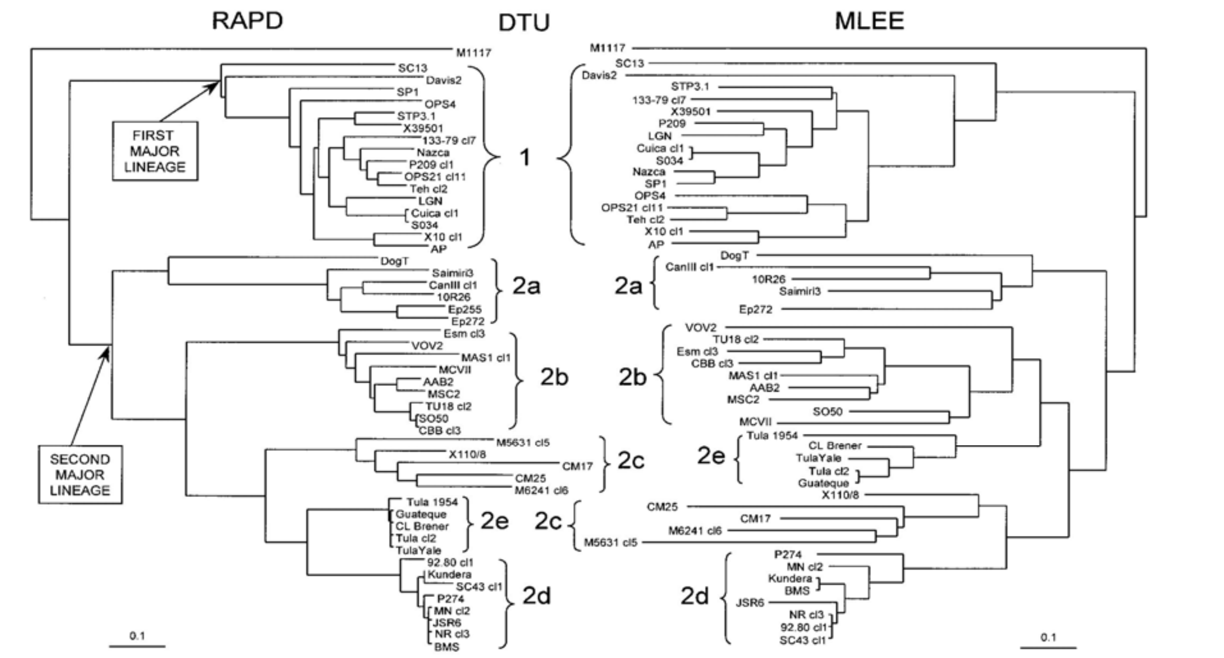


Figure 2. ***T. cruzi strains and relationships****. Neighbor joining dendrograms based on the analysis of 20 RAPD primers (left) and 22 isoenzymatic loci (right) showing the genetic relationships between 49 T. cruzi stocks and T. cruzi marinkellei stock M1117. The scale indicates the Jaccard distance along the branches. Six genetic clusters were distinguished and their names are given in the central column between the dendrograms. DTU 1 corresponds to the first major lineage of T. cruzi, while the second major lineage is subdivided into DTUs 2a-e. Adapted from Brisse, Barnabé, and Tibayrenc, M., 2000.*

sequence, SL RNA sequences, RAPD analysis, isoenzyme analysis, etc.). TCI is more commonly associated with the sylvatic cycle and is considered to cause relatively innocuous or asymptomatic infections in humans. TCII species are more likely to cause acute and chronic disease in humans, in particular in domestic environments of the Southern Cone of South America. There is some correlation between the subgroup (a-e) of the Group II isolates and different clinical manifestations (digestive tract, cardiac, or asymptomatic) in human infections. Some isolates (e.g., those of TC IIb) are highly virulent and lethal for mice, whereas others (TC I; TC IIc) seem to induce largely asymptomatic long term infections. The genetic divergence between different subgroups of TCII is about 1% (SSU rRNA sequence). The CL Brener strain that was sequenced recently is a member of Group IIe. The condition of the assembly of the genome of this isolate is less than optimal. Thus, additional genomic sequences of *T. cruzi* isolates are sorely needed.

In addition to the above, several additional criteria are helpful in selection of *T. cruzi* isolates for sequencing. First, the ‘type’ strains for some of the *T. cruzi* groups are available. These strains have been studied and the availability of this data will be essential for correlations of genotypes with phenotypes. Second, isolates that have characteristics (clinical manifestations, pathogenesis, etc.) that do not conform to their ‘genetic’ group are available in some cases, providing a unique opportunity to identify the genetic bases of these characteristics. Other factors such as host and vector association, geographic origin, and level of heterozygosity will also impact strain selection. Based on these criteria, we suggest the following isolates for comprehensive genomic analysis (see table).

Finally, attempts will be made to use isolates that are as close as possible to 'native' in order to ensure that the characteristics of the human pathogenesis are preserved. Moreover, pathogenesis in murine models and cell culture will be confirmed. However, as described above, it is most important at this stage to generate reference sequences for each of the major subtypes to which other isolates can eventually be compared. Thus, the appropriate strains will be selected with those criteria in mind.

Lineage	**Strain**	**Geographical origin**	**Host**	**Clinical/Profile**	**Hybrid**	**Priority**	**Avail**
TCI	Silvio X10	Brazil/Para	Human	Acute Phase	-	1	Yes
TCI	Jose	Brazil/Paraiba	Human	Chronic /Card Severe	-	2	Yes
TCIIa	Can III	Brazil/Para	Human	Chronic/asymptomatic	-	1	Yes
TCIIa	Tc11	Brazil/Para	Aotus	Asymptomatic	-	2	Yes
TCIIb	Esmeraldo	Brazil/Bahia	Human	Acute Phase	-	1	Yes
TCIIb	Tc872	Brazil/Goias	Human	Chronic/Digest Severe	-	3	Yes
TCIIc	3869	Brazil/Para	Human	Acute Phase	+?	1	Yes
TCIId	NRcl3	Chile/Salvador	Human	Chronic/Indeterminate	+	1	Yes
TCIId	Tc656	Bolivia	Human	Congenital	+	3	Yes
TCIIe	Tula cl2	Chile	Human	ND	+	1	Yes
TCIIe	CL Brener	Brazil/Rio Grnde Sul	insect	NA	+	done	Yes

Lineage I isolates. Lineage I (TCI) isolates, which seem to be more homogeneous than TCII isolates, have traditionally been considered to be primarily of the sylvatic cycle and are often associated with marsupials (*Didelphis*) and triatomine bugs of genus *Rhodnius* inhabiting arboreal ecotopes. They are frequently considered to be less pathogenic for humans than Lineage II (TCII) isolates. There is also evidence that the route and mechanism of cellular invasion of TCI isolates differs from that of TCII isolates in cultured cells, as well as the virulence and tissue tropism in murine models. Thus, examination of TCI isolates has the potential to identify the genetic bases of these differences.

Silvio X10/4 is a well-characterized isolate from the acute phase of a human infection in Belem, Para, northern Brazil. As for other Lineage I isolates, this isolate induces a very low parasitemia and causes chronic infection in mice. It should be sequenced as the typical 'type strain' for TCI. Its pathogenesis in murine models and cell culture has been well characterized. This strain is maintained in the labs of Michael Miles and in the TCC maintained by Marta Teixeira and have been carefully cultivated for verification of pathogenicity. It represents a clear type strain for this group and therefore was selected for high priority sequencing.

Jose is also a well characterized TCI isolate from Northeastern Brazil, but it was isolated directly from infected cardiac tissue from a Chagasic patient and thus shows the unusual (for a TCI) characteristic that it causes severe chronic cardiac disease. This is a recent isolate that has been preserved in the Trypanosome Culture Collection in its native state without significant laboratory passage.

Examination and comparison of **Silvio X10/4** and **Jose** could identify the reasons for the phenotypic differences. Other Group I isolates that were mentioned and eventually need to be sequenced include Brasil, Dm28, G, CA1, etc.

There was a general consensus among investigators that Silvio X10 was the most relevant Lineage I isolate to be sequenced. Of all the *T. cruzi* isolates, it is the most important.

Lineage II isolates. Sympomatic and severe human Chagas Disease is most commonly associated with TCII isolates, and TCII is most commonly associated with the domestic cycle of the disease. TCII isolates are more heterogeneous than TCI isolates, and have been partitioned into five groups: TCIIa-e. Evidence is accumulating [see fig. 2] that TCII emerges with two phylogenetic clades (a-b), with three apparently hybrid lineages (c, d, e). There are interesting hypotheses concerning the origin of the hybrid lineages, and this information is likely to be provided by their genome sequences.

CanIII and Tc11 (TCIIa). TCIIa isolates, originally classified as Z3B, are widely distributed in the Amazon region, Venezuela, Colombia and, probably even in the US. This group is predominantly sylvatic and, despite its broad distribution, infections in humans are rare, with most cases reported in the Brazilian Amazon region. Can III was isolated from chronic asymptomatic patient and is one of the best-studied representatives of TCIIa strains, i.e., a type strain. It is apparently not a hybrid and therefore represents an ideal sequencing target. There was good consensus within the *T. cruzi* research community that Can III is the preferred isolate of Lineage IIa for sequencing. A second TCIIa isolate of great interest is Tc11. This isolate has also been well characterized. It is of particular interest as an isolate that is closely related to CanIII, but that was isolated from a squirrel monkey (*Aotus* sp.) – an very unusual host for *T. cruzi*.. Thus, comparison of CanIII and Tc11, would not only confirm the genetic makeup and characteristics of TCIIa, but would potentially identify differences associated with the ability to infect *Aotus*. The pathogenicity of CanIII is well-established in model animals and Tc11 represents a related recent isolate.

Esmeraldo and Tc872 (TCIIb). TCIIb, like TCIId and IIe, seem to be confined to the southern cone countries. Isoalates of TCIIb are agents of more severe Chagas disease and circulate primarily domestic transmission cycles, and are transmitted by *Triatoma infestans*. TCIIb isolates, like TCI strains, appear to be less complex than other groups. Murine infections with IIb isolates exhibit high parasitemias and other differential pathogenic effects – including high mortality - when compared to those of other lineage II subgroups. Esmeraldo, which was isolated from an acute phase patient in Bahia, Brazil, has already been partially sequenced. There is general consensus that the sequence should be completed. Isolate Tc872, also ascribed to lineage TCIIb, was isolated from a chronic phase patient in Goias, Brazil. This patient exhibited digestive manifestations and comparison of data from it to data from other lineages would provide interesting and valuable information which could identify the genetic characteristics associated with the digestive form of Chagas disease. The pathogenic characteristics of Esmeraldo have been well documented and are stable, and the Tc872 represents a related, recently isolated strain.

3869 (TCIIc). TCIIc isolates are widely distributed in South America and, eventually, also in the US. Isolates of this lineage circulate in the sylvatic cycle associated with vectors and mammals inhabiting terrestrial and semi-fossorial habitats. TCIIc isolates are apparently hybrids with some relationship with TCI and TCII isolates. 3869 is a well-studied TCIIc strain isolated from an acute phase patient. Sequencing of 3869 will provide insights into how these hybrids formed in addition to the genetic basis of their pathogenic potential.

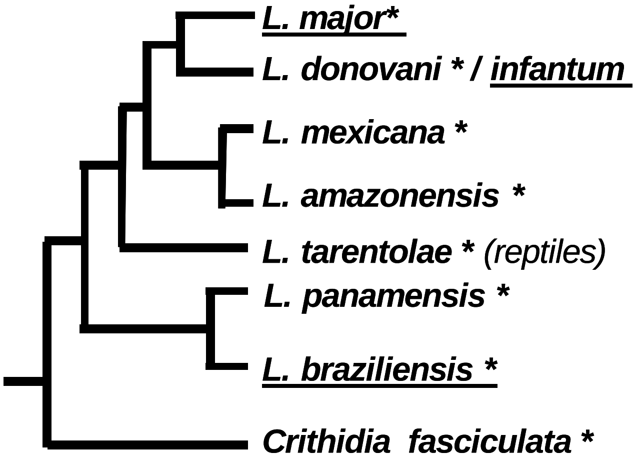
3869 is closely related genetically to other TCIIc isolates. It was recently isolated and has not been passaged significantly in the lab.

NRcl3 and Tc656 (TCIId). TCIId and TCIIe strains include hybrid strains that have haplotypes apparently split across the IIb and IIc groups. Like TCIIb and IIe, TCIId is apparently associated with the domestic cycles of transmission, and are confined to the southern cone countries where they cause serious disease. TCIId is a very interesting lineage because it remains unchanged over a broad geographical range (Chile, Brazil, Argentina and Bolivia) and maintains heterozygosity at several loci suggesting that natural selection has favored this unusual genotypic combination. Murine infections with IId isolates do not yield high acute phase parasitemias, but seem to exhibit more significant chronic phase parasitemias than do IIb isolates. NRcl3 is the reference strain of this subgroup and was isolated from a Chilean patient with chronic Chagas disease. Its pathogenic characteristics in murine systems is well-established and maintained. Another interesting isolate is Tc656, which was isolated from a very rare case of congenitally acquired Chagas disease in Bolivia. Whereas congenital transmission is rare in most geographic regions, in regions where these isolates are found, this type of transmission is observed. Tc656 has not been extensively passed in the lab.

Tula cl2 (TCIIe). CL Brener, the *T. cruzi* isolate that has been most extensively sequenced, is a TCIIe isolate. Along with lineage IId, TCIIe isolates are the predominant strains transmitted in the domestic cycle in Brazil, Bolivia, Chile, Argentina and Paraguay. They are hybrid isolates, and the hybridization has been proposed to have occurred in the sylvatic transmission cycles. Since CL Brener was isolated from an insect vector, Tula cl2, which was isolated from a human infection, would provide an interesting comparison to the existing sequence. In addition, the Tulahuen isolates have been extensively studied since they were first isolated in Chile several decades ago. Several investigators also suggested that the RA isolate, which causes a virulent and lethal infection in mice presents a good alternative. The consensus was to sequence a well characterized Tulahuen isolate. Tula cl2 has been well-characterized in terms of its pathogenesis and, although it has been maintained in the lab for a significant length of time, efforts will be made to use an isolate with appropriate pathogenic characteristics.

C. The *Leishmania*. The genus *Leishmania* comprises several subgenera and species complexes. As in the other trypanosomatids, molecular methods have proven central to the identification and classification of the various lineages. Within each group several species have been identified as pathogenic to humans, typically with distinctive pathologies although there is considerable overlap. The molecular basis for these, especially in conjunction with variation in the human host, is largely unknown and ripe for pursuit by comparative genome sequencing. The properties of the individual *Leishmania* under consideration and the sequencing rationale and strategies appear in following sections. As outlined earlier, our preferred strategy involves sequencing of key lineages as several closely related strains/species within these. SSU rRNA comparisons show that *Leishmania* species typically diverge by less than 1%, and show about 2% difference from the insect trypanosomatid *Crithidia fasciculata*.

Current status. To date a complete finished genome has been generated for a member of the *L. tropica* species complex (*L. major* strain Friedlin), *L. donovani* complex (*L. infantum* JPCM5) and *L. braziliensis* complex (cutaneous *L. braziliensis* M2904). The *L. major* sequence was generated by a consortium of laboratories and is considered finished. *L. infantum* and *L. braziliensis* were completed at the Sanger Institute; assemblies were generated after ~5X sequencing and scaffolding vs. the *L. major* assembly, with limited confirmatory sequencing and or finishing.



Several sequencing initiatives are underway at the Sanger Institute. An isolate of the *L. mexicana* complex (*L. mexicana*) has been selected and is currently in the sequencing pipeline. Two isolates of *L. donovani* are being identified. In the figure above, species for which complete genomes are available are underlined, while additional species proposed here or already in progress are marked by an asterisk.

Leishmania	*Strain*	*Geographic origin*	Clinical profile	Priority	Avail.
L. major	MHOM/IL/80/Friedlin	Israel	Cutaneous (healing)	Done	N/A
	Seidman	West Africa	Cutaneous (nonhealing)	1	Yes
	LV39 (Neal) or Iran	Southern Russia/Uzbekistan	cutaneous	2	Yes
L. tropica	TBD		Cutaneous	1	Yes
L. aethiopica	TBD	Ethiopia	Cutaneous	2	Yes
L. donovani	TBD*	Africa, India	visceral	1*	TBD*
L. infantum	MCAN/ES/98/LLM-877	Mediterranean basin	asymptomatic, visceral	Done	N/A
L. mexicana	MHOM/GT/2001/U1103 *	S/Central America	cutaneous	1*	N/A
L. amazonensis	TBD	S/Central America	cutaneous	2	Yes
L. braziliensis (mucocutaneous)	TBD – 2 strains cutaneous, 2 strains mucocutaneous	South America	mucocutaneous	1	TBD
L. braziliensis (cutaneous)	MHOM/BR/75/M2904	South America	cutaneous	Done	N/A
L. panamensis	TBD	S/Central America	cutaneous	1-2	Yes
L. tarentolae	Parrot-II	Old world	Lizard parasite	1-2	Yes
Endotrypanum	TBD	S,/Central America	Sloths	2	Yes

***** these strains have been targeted for sequencing at the WTSI.

***L. tropica* complex.** The most finished genome of any trypanosomatid currently is that of *L. major* Friedlin, which is essentially completed.

a) For *L. major*, a key deficiency is the unavailability of significant numbers of cDNA/EST sequences. These are critical for a variety of analyses, chief of which is transcript definition, in order to characterize *cis* acting signals used in processing of polycistronic mRNA precursors, and to define 5' and 3' UTRs important in the regulation of mRNA abundance and translation. At present only 2-3K ESTs are available, mostly from another strain (LV39, discussed more below). Remarkably these ESTs have provided evidence for unanticipated transcript and/or gene classes not evident from protein 'ORF' based annotation efforts. Thus an important priority is generation of a sufficient number of EST sequences, from the three major life cycle stages (procyclic and metacyclic promastigotes, and lesion amastigotes). Ideally these would be generated from normalized cDNA libraries and in sufficient numbers to cover the great majority of transcripts. Since as discussed earlier, trypanosomatids show virtually no *cis*-splicing, this may be a perfect opportunity for high throughput methods focusing on just the temini, such as the 5'-RATE (Gowda et al 2007, *Nature Protocols* 2: 1622) or related methods which yield both termini simultaneously.

b) The Friedlin strain arose from a classic case of Old World cutaneous leishmaniasis. Other strains of *L. major* show geographic variation in various parameters, and yield slightly different pathologies, in humans and in some cases were tested, in animals. One example is a Western African Strain termed Seidman or Sd (MHOM/SN/74/SD); this strain arose from a non-healing lesion in a patient and in laboratory animals; it has been extensively characterized by David Sacks' group. Other isolates suggested by the community include the i) LV39cl5 line (also termed the Neal or NIH line), originally from southern Russia/Uzbekistan; this line was the source of most of the existing ESTs; and ii) the "Iran" strain used in live vaccination studies there. Well characterized isolates for all of these strains are available for this initiative.

c) The two other major lineages within the *L. tropica* complex are *L. tropica* sensu strictu and *L. aethiopica*. Of these *L. tropica* warrants the most interest, including differences in pathology and susceptibility to drugs, and its emergence as a potentially anthroponotic disease in Central and West Asia. A series of lines were described and characterized molecularly recently by Hadighi et al (PLOS Medicine 2006), and one of these would make a logical candidate although others could be considered. *L. aethiopica* infection is endemic in Ethiopia, Kenya and South-West Africa and has been found in Saudi Arabia.; this parasite has not been extensively studied. Unlike *L. major* or *L. tropica*, *L. aethiopica* can cause diffuse, mucocutaneous and limited cutaneous leishmaniasis; LCL caused by *L. aethiopica* can persist for years. Comparison of this organism to *L. major* and *L. tropica* as well as *L. amazonensis* (another species causing DCL) would provide new insights on the basis of this pathology.

***L. donovani* complex.** *L. donovani* sensu strictu is the predominant cause of fatal visceral leishmaniasis. At a molecular level, it is closely related to *L. infantum*, which however typically generates mild or even asymptomatic pathology in humans, only emerging as visceral leishmaniasis upon immunosuppression. The *L. infantum* sequence has been completed at the Sanger; *L. infantum* is widespread in the Mediterranean basin but is associated primarily with a mild disease which healthy people will resist and develop immunity. However in the face of immunosuppression, most commonly by HIV currently, infected people go on to full-blown visceral disease. *L. chagasi* represents a closely related lineage transplanted to the New World where it is associated with visceral disease as well.

Generally *L. chagasi* is considered to be closely related and even synonymous to *L. infantum* (Lukes et al *PNAS* 2007), and there was not a strong sentiment for determination of its sequence. In contrast there was universal agreement that the sequence of one or more strains of a visceral strain of *L. donovani* sensu strictu should be determined. This is presently under consideration at the Sanger Institute; it clearly warrants a high priority.

***L. mexicana* complex.** As noted earlier, *L. mexicana* has entered the sequencing 'pipeline' at the Sanger Institute. Within this species complex, several investigators advocated *L. amazonesis*. It is particularly virulent, distributed throughout South America, and it causes both limited cutaneous and diffuse cutaneous disease in humans, frequently resulting in a host that is "anergic" to infection, (i.e. displaying little/no T cell response). In mice, this parasite shows distinctive immunological responses of relevance to human disease. For example, there are no "resistant" mouse strains; all are susceptible and develop chronic disease; T regulatory cells ameliorate disease, IFNγ can enhance macrophage infection, and there is no pathology in the absence of a host immune response. Thus a comparison of these organisms to other *Leishmania* would allow identification of genes underlying its distinctive biology. Several well characterized strains are available including Marina or LTB10016.

Subgenus *Viannia* (*L. braziliensis* complex).

a) *L. braziliensis* and mucocutaneous leishmaniasis. Strains of *L. braziliensis* can give rise to cutaneous or mucocutaneous leishmaniasis, the latter being a severe, chronic disfiguring disease. Current data show that strains taken from patients with either presentation are closely related, leading many investigators to postulate host factors as the primary determinant of disease outcome. The sequenced strain of *L. braziliensis*, M2904, was taken from a patient exhibiting cutaneous disease pathology only.

There are tremendous opportunities for genome sciences in this question. First is the possibility of comparing cutaneous vs. mucocutaneous *L. braziliensis* in an effort to identify parasite factors underlying this severe pathology. To date the markers that have been used for strain comparisons are considered unlikely to play any role in the pathogenic process; genomic comparisons amongst cutaneous vs. mucocutaneous strains could well provide SNPs and/or genes that discriminate between the two outcomes. Second, there is a tremendous opportunity to simultaneously compare the genome of the parasite and the host in these two cases, for example to look at host SNP associations. While host genomes are not part of this initiative, a strong case can be made for selection of parasite strains for which the human host DNA is also available. With the increasing power and availability of human 'SNP chips' it may be possible rapidly search for human correlates of mucocutaneous disease simultaneously with those of the parasite.

A critical factor for success in this initiative will be the availability of well chosen strains. It is not yet clear to the writing group (or even those providing the suggestions) whether the existing strains are ideal for the comparisons noted above. Thus it will be necessary to assess the available strains and if necessary, undertake additional characterization. This is especially likely for those strains where provenance of human host DNA is to be performed. In this study, it seems that minimally two well chosen strains of both cutaneous and mucocutaneous disease should be studied. While it is naïve to expect that genome level comparisons of such a small number of strains will directly lead to the relevant genetic changes, they should serve to provide an array of candidate SNP or gene differences to warrant more extensive studies of a larger number of well chosen isolates in order to confirm or eliminate candidate loci. Despite the advance work required, members of the writing group strongly believed that a joint pathogen/host effort could prove extremely powerful and informative.

b) *L. braziliensis* cDNAs and 'small RNAs'. For reasons noted earlier with *L. major*, it is desirable to map transcripts via end-sequencing of ESTs. Preliminary analysis of the inter-

ORF regions between *L. braziliensis* and *L. major* suggest there is extensive divergence, to the point that it is difficult in most cases to unambiguously associate splice acceptor signals. cDNA end-sequence data would fill this gap as well as permit comparative analysis of the signals implicated in RNA processing, which is not well understood in this organism.

RNAi. Recently the genome sequence of *L. braziliensis* revealed the presence of homologs of the RNAi pathway genes Argonaute and Dicer, which are absent in the *L. major* and *L. infantum* assemblies (Peacock et al Nature Genetics 2007). Correspondingly, functional studies have shown the RNAi pathway to be inactive in the latter two species, but active in *L. braziliensis* (Beverley and Ullu laboratories, Woods Hole Kinetoplastid Molecular & Cell Biology Meeting, 2007). This has great potential towards the introduction of new functional genomic approaches in these parasites. As part of this initiative, it would make sense to characterize the 'small RNA' transcriptome from one of the *L. braziliensis* lines showing RNAi. This is a rapidly progressing field in other eukaryotes and these small RNAs may include a variety of forms arising by diverse mechanisms, and implicated in biology ranging from RNA turnover to translation to chromatin silencing. Moreover, genome and RNA comparisons of 'RNAi-positive' with closely related 'RNAi-negative' *Leishmania* may provide important insights, a type of analysis not feasible in other taxa.

As of yet, in the trypanosomatids only in *Trypanosoma brucei* has limited siRNA characterization been performed; very preliminary data suggest that *L. braziliensis* will be similar but it is early days. Thus at this time we do not advocate a particular methodology for siRNA isolation and cloning; several alternatives are widely available and at present any one of these should suffice. The source RNA could be readily generated by the Ullu or Tschudi laboratories (amongst others), who have expressed willingness to do so. The strain from which these RNAs are isolated should preferably be one of ones shown to exhibit functional RNAi pathways, such as *L. braziliensis* M2904 whose genome sequence is available.

c) *L. panamensis*. *L. panamensis* was strongly advocated by several groups, for several reasons including i) distinctive pathology in humans, ii) probable occurrence of the RNAi pathway (based upon Argonaute typing and siRNA formation data from the Beverley & Ullu groups), and iii) the availability of well characterized lines and a murine infection model. Specifically, lines developed by Nancy Saravia (Cali, Columbia) and under study in Diane McMahon-Pratt's laboratory would make logical candidates (*L. panamensis* MHOM/CO/XX/1989, or CIDEIM1989).

Subgenus *Sauroleishmania* (*L. tarentolae*).

Members of this group are pathogens of lizards rather than mammals. *L. tarentolae* has received extensive study due to its ease of laboratory culture and manipulation, which has facilitated numerous molecular and biochemical studies. Ouellette and Papadopoulou have shown that *L. tarentolae* can even infect macrophages at 37°, although it cannot divide and survive long. Thus as a model 'leishmania' this organism is highly attractive.

A consortium of the Quebec *Leishmania* groups (Ouellette, Papadopoulou, Corbeil, Tremblay and Olivier) have initiated a pilot 454 sequencing project; initially they hope to have ~100 Mb of sequence and further 454-based and/or WGS sequencing will likely ensue thereafter.

Genus *Endotrypanum.* Parasites of the genus *Endotrypanum* are closely related to *Leishmania*, by molecular comparisons diverging from *Leishmania* just after their separation for the 'insect' trypanosomatid ancestor. *Endotrypanum* is widespread in South and Central America, although it is not considered a human pathogen. One of its notable features is its residence within red blood cells, a feature unique in the trypanosomatid world. Many investigators felt this would be an exciting parasite to study for many reasons, including its novel biology. A number of strains are available. Due to the limited research community and some reservations about the value of this information to human disease, these studies were not given the highest priority.

D. Outgroups.

Selectively informative outgroups that are closely related to the human pathogens, must be studied to understand the genetic nature and evolution of parasitism/pathogenesis. Each of the major pathogenic groups is flanked by related species that, though parasitic, are apparently non-pathogenic for humans (outlined below).

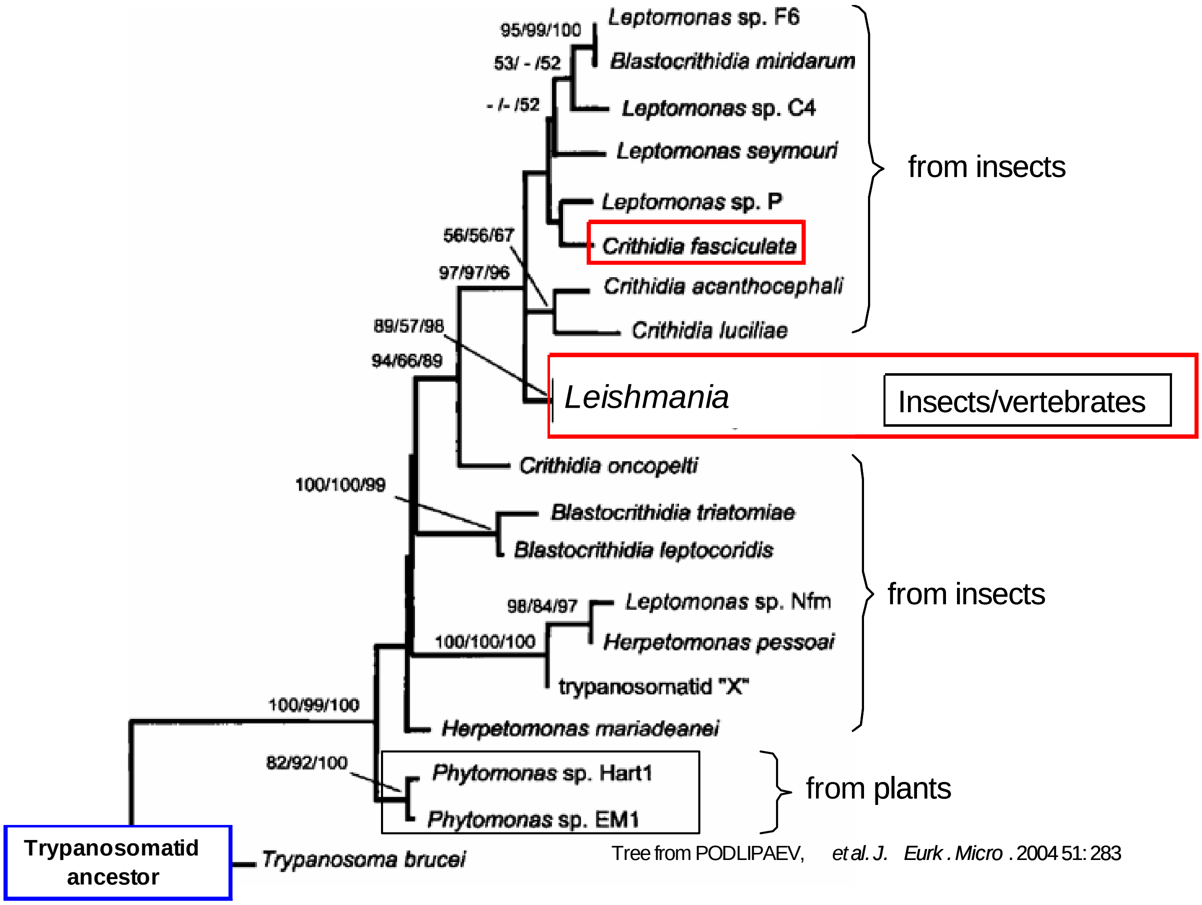
Lineage	Strain	Host	Outgroup for	comments	Priority
T. rangeli	AM80 (Group B)	Human	American trypanosomes	Salivarian trans	1
Crithidia fasciculata	Cf-C1	Insect	*Leishmania* (close)		1
C. deanei		Insect	*Leishmania* (more distant)	Has endosymbionts	3
Bodo saltans		Free-living	All pathogenic tryps	Closest bodonid	3

Clade *T. rangeli*. *T. rangeli* is generally considered to be most closely related to the American trypanosomes (it is generally considered an American trypanosome), while also exhibiting some characteristics of the African trypanosomes (see fig. 1). In fact, *T. rangeli* is the only trypanosome besides *T. cruzi* and the *T. brucei* clade that infect people. Moreover, *T. rangeli* is the only member of that group that is not pathogenic in people. The SSU rRNA of *T. rangeli* isolates differ from those of *T. cruzi* or *T. b. brucei* isolates by approximately 8% and 20%, respectively. Its geographic distribution and host range overlap that of *T. cruzi*, it is spread among susceptible mammalian hosts by species of reduviid bugs, and its sequences suggest a clustering with *T. cruzi*. However, *T. rangeli* isolates cause no known pathology in infected human or other mammalian hosts. This species is transmitted only by triatomine bugs of the genus *Rhodnius*, following the salivarian route since the parasite migrates to and replicates within the insect salivary glands. However, although infective forms are found in the salivary glands leading to salivarian transmission, the contaminative route of infection is also possible. Moreover, *T. rangeli* is apparently exclusively extracellular, in contrast to *T. cruzi* which must replicate inside cells in the mammalian host. Thus, *T. rangeli* presents a very interesting 'out-group' to the various *T. cruzi* lineages, as well as to the African trypanosomes. Questions relating to pathogenesis, intracellular invasion, salivarian transmission, etc. will be addressed by examination of the genome of this organism.

T. rangeli isolates are now generally grouped into four major lineages, A-D. Whereas, ideally, one member of each group should be sequenced, there is already a significant databank of cDNA sequences available from Edmundo Grisard's group in Santa Catarina, Brazil. These two strains, SC-48 and Choaci, represent two of the four major lineages (D and A) and are closely phylogenetically. Lineage C, which predominates in Central America, is also more related to lineages A and D than to lineage B. Thus, lineage B is the most divergent among the four *T.*

rangeli lineages. Taking advantage of the availability of the sequences of SC-48 and Choaci, and the consideration that it is unlikely that NIH will elect to sequence more than one of these non-pathogens; we propose to sequence isolate AM80, from *T. rangeli* lineage B. AM80 was isolated from an asymptomatic human patient living in the Brazilian Amazon. Comparison of the genome of this isolate with the sequences from the other *T. cruzi* isolates (as well as the sequences available from Santa Catarina) will provide a framework for understanding the pathogenesis of *T. cruzi*.

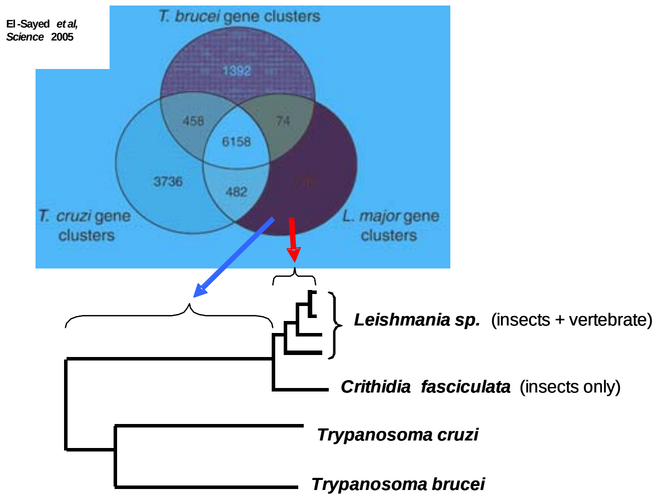
Insect trypanosomatids: *Crithidia* as the nonpathogenic outgroup to *Leishmania*. Following the ancient divergence from the trypanosome lineage, the lineage leading to *Leishmania* shows a series of divergent lineages leading to groups such as *Phytomonas*, *Herpetomonas*, *Leptomonas*, *Blastocrithidia* and *Crithidia.*



Tree from PODLIPAEV, *et al. J. Eurk. Micro.* 2004 51: 283

<u>*Crithidia fasciculata.*</u> Several lineages of 'insect trypanosomatids' diverge immediately prior to the origin of the pathogenic *Leishmania* and its closely related genus *Endotrypanum*. As such they represent logical candidate 'outgroups' for the origin of vertebrate parasitism in the *Leishmania*. Examples of these include *Crithidia fasciculata*. Development takes place only in the guts of insects, and transmission is by contamination with feces or insect ingestion by predator insects. These trypanosomatids are not restricted to only one host, and can be harbored by flies, mosquitoes and hemipterans (hematophagous, predators and phytophagous). *Crithidia* can be cultivated in high yields using inexpensive undefined media or fully defined serum-free media and does not require specific bio-safety precautions. Since the basic cellular machinery of *Crithidia* is similar to that of the other pathogenic trypanosomatids, it is an excellent model

organism for biochemical and molecular studies. Moreover, as representative of one of the closest 'outgroup' lineages to *Leishmania*, *Crithidia fasciculata* is ideally placed in order to identify genes specifically involved in the acquisition of vertebrate parasitism from the ancestral insect-only parasite.



For *C. fasciculata*, a pilot sequencing project of the Cfc-1 line obtained from Larry Simpson was initiated by the Beverley laboratory. Thus far 8-10X coverage has been obtained by 1st-generation 454 sequencing, as well as paired end sequences from 30,000 WGS shotgun clones. Preliminary assemblies and comparisons have been encouraging but suggest that additional sequencing will be required to get a reasonable coverage by either scaffolding or other approaches. These could build on the existing data set cooperatively. There is also some suggestion that the degree of heterozygosity evident in these assemblies is higher than seen previously in other trypanosomatids. Cf-C1 was originally provided by William Trager at Rockefeller in the 1960's.

A variety of other interesting insect or plant trypanosomatids are evident in the evolutionary tree leading to *Leishmania* and C*rithidia fasciculata*. While of great interest evolutionarily, we have not prioritized these for study here. Some work focusing on the plant trypanosomatid *Phytomonas* and the endosymbiont bearing *Crithidia deanii* have been started elsewhere.

The Bodonids: a non-parastic, nonpathogenic outgroup of trypanosomatid parasites. All trypanosomatids are parasites of plants, vertebrates or invertebrates. Thus examination of their closest free-living protistan relatives provides some opportunity for probing the origins of parasitism globally. Of these, members of the *Bodonina* present the best opportunities, including several species in the families, *Bodonidae* and *Cryptobiidae*. One negative is the fact that these are distant relatives, showing SSU rRNA divergences of 18% or more. Such evolutionary distances may hinder one's ability to precisely associate genetic changes with phenotype. While there are various approaches to overcome this problem, and many on the working group judged this to be of great importance, ultimately we gave this low priority (3). Notably the Bodonid that probably represents the closest outgroup, *Bodo saltans*, has been targeted for sequencing at the WTSI. J. Lukes (Czech Republic) has also expressed willingness to generate genomic DNA for this purpose should it be necessary.

III. Sequencing Strategy.

Trypanosomatid genomes are uniquely amenable to large scale sequencing efforts for several reasons. First, the genomes are relatively small, typically 25-35 Mb. Second, even across the three major lineages, a high degree of synteny is observed. This synteny is helpful in both the sequence assembly and annotation phases. Moreover, its power grows as the evolutionary distances decrease, as seen in comparison of the *L. braziliensis* and *L. infantum* genomes (Peacock et al, Nature Genetics 2007). Third, genes are organized in polycistronic arrays and lack introns, further simplifying gene identification. Fourth, while the genomes are typically diploid, heterozygosity, where examined, has been relatively low (<0.01%). There can also be challenges in sequencing trypanosomatid genomes. The genomes typically contain numerous extensive gene families of varying degrees of homogeneity, many of which are thought to play important roles in the infectious cycle, which can present challenges in assembly (e.g., *T. cruzi* CL Brener). Similarly, transposable elements, which can also impact the assembly process, are numerous in some genomes (e.g., *T. brucei* and *T. cruzi*).

A. Complete, finished genome sequence or high quality draft coverage? Whereas completed high quality linear sequence of each of the chromosomes of each of the organisms to be studied is the 'holy grail' of the sequencing effort, this may be not be necessary for the majority of the needs of trypanosomatid researchers. However, a very significant loss if the genome is not completed is the ability to dissect the genome architecture, and therefore the identification of potential genome rearrangements, transpositions or deletions associated with pathogenic or other phenotypes of the organisms. Thus, it is desirable that the genomes be sequenced and assembled into end-to-end chromosome size contigs as much as is possible.

Where examined, and most of these parasite genomes have been characterized by PFGE, the genome size estimates range from ~25 -- 50 Mb. Trypanosomatids are minimally diploid, but in some cases evidence for aneuploidy and/or polyploidy has been obtained. Thus, there is a significant opportunity for the generation of SNPs that will have many applications in molecular epidemiology and population genetics, an important output for this project. Moreover, in *T. cruzi*, several lineages are hybrids (e.g., the CL Brener line that was the object of previous sequencing efforts). These pose both challenges and opportunities and are discussed below.

Whereas the choice of specific technologies will be made by the sequencing centers, the writing group was of the opinion that in general it is important to have a high degree of coverage/ confidence in the sequence assemblies. The substandard assembly of the previously sequenced CL Brener isolate of *T. cruzi* provides an example of the necessity for high quality data. To date, it has been impossible to achieve a high quality sequence from this data. Moreover, both the Buck and Beverley labs have had the opportunity to evaluate 454 sequencing technologies in trypanosomatids. This experience suggests that the degree of coverage required for confident assemblies is high; probably greater than 20X, as currently recommended by the manufacturer. Current 454 assemblers have particular difficulty with intergenic regions, which are replete with simple sequence repeats in trypanosomatids, and the short 454 reads, as expected, face significant challenges with repetitive gene families. In combination, these difficulties give an incorrect appearance of frequent 'syntenic breaks'. Thus, it appears that some hybrid strategy (for example 454 combined with paired end sequencing of cloned DNA) will be the most efficient route for final sequencing of these genomes.

The average %GC should not present a problem since the known base composition ranges between 40 and 60%. Repetitive sequences (gene families as well as transposable elements) may be an issue, as they were with the analysis of the first *T. cruzi* genome (strain CL Brener). As also seen with this strain, several of the high priority isolates of *T. cruzi* are also likely to be hybrids. Although this fact may complicate assemblies, analysis of the non-hybrid isolates will be helpful in dissecting these genomes. The origins of these hybrids is in fact one of the important outstanding questions that can be addressed by this analysis.

For selected key species (see Table 1, page 2), a finished sequence by whatever means necessary (equivalent to at least 10X coverage by traditional shotgun sequencing, gap closure by PCR amplification and sequencing) is highly desirable. Similarly, for some of these key species, a deep EST library would be helpful. These strains were selected for sequencing because they are representatives of 'type strains' of each of the major subgroups of the three major lineages of pathogenic trypanosomatids. These are discussed below:

a. African trypanosomes. As described above, *T. b. brucei* Lister 427 represents the major subgroup of *T. brucei* that is not pathogenic for man. *T. b. rhodesiense* Uganda represents the only member of this subgroup being sequenced. *T. evansi* is apparently not infectious for humans, is the only 'African' trypanosome found outside of Africa, and is spread mechanically rather than from the salivary gland of a tsetse fly. Comparison of these sequence to that of *T. b. brucei* and the *T. b. gambiense* being sequenced at the Sanger Center will provide insight into the major pathogenic differences these subspecies exhibit (see above). We believe that a very high quality draft sequence analysis is likely to provide this information. Since one isolate of *T. b. brucei* 927 has already been sequenced, these drafts can be compared to the existing sequence of 927 to identify differences. We therefore recommend closure of *T. evansi*.

b. American trypanosomes. We recommend closing the sequence of *T. cruzi* isolates Silvio X10 and Can III. These represent one member of the TCI lineage of *T. cruzi*, and one member of TCII, which are some of the most common and best studied. These two groups are also considered to be diploid, whereas TCII c, d and e, are considered to be 'hybrids' containing more than a diploid number of chromosomes. The latter will also be more challenging to close; i.e., the *T. cruzi* CL Brener isolate that was sequenced previously presented extreme challenges because of its hybrid nature (it is a member of TCIIe). Thus, closure of Silvio X10 and Can III will provide a format for base by base comparisons of the large phenotypic and pathogenicity differences between the TCI and TCII lineages of *T. cruzi* (the SSU rRNA sequences of TCI and TCII isolates vary by about the same amount as those of *Leishmaniai* and *Crithidia*), and provide insight into the differences between the subgroups of the TCII lineage. High coverage draft sequence of strains Esmeraldo, 3869, NR cl3, and Tula will provide a basis for identifying the differences between the TCII b, c, d and e sublineages, without introducing the great complexity of completing these sequences.

c. *Leishmania*. Three genomes of *Leishmania* have been finished and work on *L. mexicana* has been started. The writing group felt that completely finishing the genome of the additional species prioritized here would be unnecessary, given their relatively close relationship.

d. Outgroups. The two 'out groups' we have recommended with highest priority, *T. rangeli* and *C. fasciculata*, should be sequenced to completion. *T. rangeli* varies from the other American trypanosomes by approximately 8% in the sequence of it SSU rRNA (~20% from the African trypanosomes). Its complete sequence will provide a very good basis for genome comparisons between the American and African trypanosomes. Thus, it is recommended that the genome be sequenced to completion. Similarly, *Crithidia fasciculata* is the ideal outgroup for the *Leishmania*. Other *Crithidia*, e.g., *C. oncopelti* and *C. deanei*, are more distant from the *Leishmania.*

The complete sequences of all of the strains discussed will provide a reference to which all the genomes, completely sequenced or ‘in draft’, can be compared.

B. Transcript mapping by EST/cDNA sequencing. The writing group was of the opinion that it is more important to sequence the genomes from more organisms rather than to generate cDNA sequence data from each species, for reasons evident below. However for the purpose of transcript mapping, cDNA sequencing was recommended for only a strategic set of species.

In trypanosomatids mRNAs are formed by processing of a polycistronic precursor RNA into monocistronic mRNAs bearing a 5' capped 39 nt ‘miniexon’ added by *trans*-splicing, and a conventional poly-A tail. Only three instances of conventional *cis*-splicing have been reported (Mair et al 2000; RNA 6:163). Thus in effect, only mapping of 5' and 3' cDNA ends are required to characterize typical mRNAs. This lends itself well to newer 454 based methods.

One major value of cDNA sequencing is first in transcript unit definition, including 5' and especially 3' UTRs implicated in the control of translation and mRNA abundance. This information in turn can be associated with mRNA abundance data arising from expression profiling studies in order to detect regulatory sequences. These cDNA sequences will also provide information that is likely to permit identification of the appropriate translational start and stop sites, and the splicing and polyadenylation signals. The lack of this information has been a major impediment in the analysis and use of the genomic sequences of the three existing trypanosomatid genome sequences. Finally, this cDNA sequencing is essential to find genes that are not found by informatics approaches (experience in the Buck lab with a cDNA library from *T. dionisii* suggests that over 10% of the cDNAs are not represented in the libraries of genes predicted by gene finders in *T. cruzi*, *T. brucei*, *Leishmania*, or *T. dionisii* itself). In the Table on page 2, we identify a small subset of strategically selected isolates for which we recommend relatively deep sequencing of normalized cDNA libraries. In our experience (cDNA library of *T. dioinisii*, Buck lab unpublished), analysis of 20,000 cDNA clones from both ends has provided sequences of nearly 7,000 discrete genes. We therefore recommend at least this level of sequencing for each of the strains selected.

a. African trypanosomes. We have recommended deep sequencing of a normalized cDNA library for *T. b. brucei* and *T. evansi*, representing the two major groups of parasites (e.g., those pathogenic for man and transmitted by tsetse fly -- *T. brucei*; and those not pathogenic for man and mechanically transmitted – *T. evansi*). We have not included *T. rhodesiense* in the group for generation of EST libraries, because it is probably similar enough to *T. brucei* that little additional information will be gained.

b. American trypanosomes. We recommend sequencing normalized cDNA libraries from *T. cruzi* Silvio X10 and Can III to provide a baseline for annotating and comparing the genomes of the TCI and TCII lineages of *T. cruzi*. We recommend performing a cDNA analysis of only one of each of the major subgroups as we believe that the isolates within each subgroup will be served nearly equally by the sequences from a single isolate.

c. *Leishmania*. We recommend end sequencing from normalized cDNA libraries from two strains: *L. major* Friedlin and *L. braziliensis* M2904. Since *L. braziliensis* (unlike the ‘higher’ *Leishmania*) has active RNAi machinery, it is also recommended that small RNA libraries of *L. braziliensis* be generated and sequenced deeply.

d. Outgroups. We recommend sequencing cDNA libraries from *T. rangeli* and *C. fasciculata*. It is unlikely that the cDNAs from any of the other isolates will be sufficiently similar to those of these two organisms to assist in transcript mapping and annotation. These are the optimal outgroups for the American and African trypanosomes and the *Leishmania*.

C. DNA purification. Strains selected will be cloned where possible and grown in sufficient quantity in vitro. DNA will be isolated by standard approaches. As much as 50% of the DNA content of some of these organisms is kDNA, which is located in the kinetoplasts, or mitochondria. The kDNA consists of tens of identical ~25 kb maxicircles encoding mitochondrial genes, and thousands of heterogeneous 1-2 kb minicircles which encode the novel guide RNAs. Although the sequence of the kDNA is of great interest, steps will be taken to remove the bulk of the kDNA from the total genomic DNA preparations to minimize redundant kDNA sequencing. This can be accomplished by banding in CsCl / ethidium bromide density gradients, size selection of sheared genomic DNA prior to cloning, or gel electrophoresis.

IV. Data repositories, curation and informatics.

Currently the primary data repository for trypanosomatid genomes is managed by the Sanger Institute (www.genedb.org). This Center has undertaken extensive curation of the 5 completed trypanosomatid genomes as well as other genomes in progress there. Other smaller/more focused repositories include CruziDB (http://tcruzidb.org), the South African National Bioinformatics Institute (www.sanbi.ac.za), and BioWebDB in South America (http://www.biowebdb.org).

There was universal agreement in the community that to accommodate the anticipated expansion of trypanosomatids genomes, every aspect of the data infrastructure needs strengthening and additional support. In our discussions two recommendations were made repeatedly.

First, it was widely agreed that the availability of centralized database repositories accompanied by high quality curation was essential for progress. These should include the ability to incorporate diverse forms of data including expression arrays, proteomics, and other datasets.

Secondly, that this information should be made available in a form suitable for relatively 'seamless' input into advanced bioinformatic tools that could be readily applied by researchers in their own work. Two basic models were advanced: a 'central' resource, perhaps along the lines of PlasmoDB or ApiDB, with a series of pre-configured powerful tools; and a 'distributed' model, where individual groups would develop their own unique analytical tools but would benefit from the standards and database schema incorporated in the centralized model.

Several investigators have expressed interest in working to implement and build upon the existing bioinformatic infrastructure necessary to support these studies, including Jessica Kissinger, Matt Berriman, and Greg Buck. Kissinger has a strong record with the various apicomplexan DBs, and Berriman has a strong record with the GeneDB-based curation. We emphasize that these and all investigators consulted stressed that funding will be required to meet the needs, both current and anticipated. The research community is discussing plans that will allow us to tackle this challenge.

The community endorses the existing and standard data release policy for these projects as outlined at http://www.genome.gov/10000925. Raw and processed/curated data will be deposited with NCBI, as for other genomes, and in GeneDB (www.genedb.org) or CruziDB (http://tcruzidb.org), as the case warrants and resources are available.

V. The community.

As part of this white paper, input was solicited from the trypanosomatid community via several routes, including a discussion at the Kinetoplastid Molecular and Cell Biology Meeting held at Woods Hole MA in 2007. Over 100 investigators at the meeting expressed their interest in the analysis of additional isolates of the trypanosomatids. Subsequently, several hundred investigators on various trypanosomatid list serves and email lists were contacted by Greg Buck, Steve Beverley, Larry Simpson, John Donelson, George Cross, and others preparing the white paper. These emails emphasized the inclusiveness of the project, and encouraged recipients to: 1) provide input from their experience and knowledge base; and 2) to contact any other investigators for their input and knowledge base, with the goal of identifying the best panel of isolates for sequencing. Many dozens of responses were generated by these requests for information. Responses came from investigators in North America, Europe, and Latin America.

In particular, responses for the African trypanosomes came primarily from North America (US and Canada) and Europe (England, France, and Germany). Responses for the American trypanosomes included investigators in the U.S. and many Latin American Countries (Brazil, Argentina, etc.), and Europe (England, France). Responses for the Leishmania were generated from the U.S., Latin America (Brazil, Columbia, Argentina), Europe (England, Germany). Thus, there is very strong, global support for the generation of new sequences for these organisms.

There were many, many responses to our requests for information. Many investigators made impassioned arguments in favor of their favorite isolates and strains. Each of these suggestions was carefully considered. In the end, there was excellent consensus about the recommendations submitted above. Perhaps most importantly, there was tremendous input from the laboratories that are most likely to be involved in various phases of the sequencing proposal including strain provision and characterization, as well as the interpretation and application of the forthcoming sequence data. The prioritization indicated in the discussion above reflects this consensus, fully acknowledging that this 'batch' of sequences will provide a basis for, but in now way completely satisfy, the sequencing appetite/requirements of the large community of investigators studying the biology and pathogenesis of the Kinetoplastida.

Investigators who have endorsed, supported and/or provided input to this proposal.

In addition to the authors of this document, the following individuals have contributed input of ideas and opinions: Diane McMahon-Pratt (Yale), Jeffrey Shaw (Brasil), Nancy Saravia (Columbia), Angela Hampshire Lopes (Brazil), Mariano Levine (Argentina), Carlos Frasch (Argentina), Michael Miles (London), George Cross (New York), Dave Barry (Glasgow), Christine Clayton (Heidelberg), Mary Wilson (Iowa), Julius Lukes (Prague), Elisabetta Ullu (Yale), Rick Tarleton (Georgia), Edmundo Grisard (Brasil), Paul Bates (Liverpool), Alberto Davila (Brasil), John Kelly (London), Mark Carrington (Cambridge), Marc Ouellette (Quebec), David Engman (Chicago), David Campbell (UCLA), John Donelson (Iowa), Larry Simpson (UCLA), Alberto Davila (FIOCRUZ), Dmitri Maslov (UC-Riverside), Marco Krieger (IBMP-Brasil), Samuel Goldenberg (IBMP-Brasil), Paul Bates (Liverpool), Mario Steindel (UFSC-Brasil), Stenio Fragoso (IBMP-Brasil), Ed Louis (Nottingham), Michael Lewis (LSHTM-UK), Mariono Levin (Buenos Aires), Carlos Frasch (UNSAM-Argentina), John Kelly (LSHTM-UK), S. Leguizamon and M.R. Gonzalez Cappa (Fibertel-Argentina), Michel Tibayrnc (France), Sergio Schenkman (Brasil), Nobuko Uchida (Brasil), Patricio Manque (VCU), Myrna Serrano (VCU), Shozo Ozaki (VCU), Barbara Papadopoulou (Quebec), Erney Camargo (Brazil), Martin Llewellyn (UK), Matthew Yeo (UK), Michael Gaunt (UK), Sergio Schenkman (Brazil), Nobuko Yoshida (Brazil), Franco da Silveira (Brazil).