

A case for a *Glossina* genome project

Serap Aksoy*¹, Matt Berriman², Neil Hall³, Masahira Hattori⁴, Winston Hide⁵, Michael J. Lehane⁶

¹*Yale University School of Medicine, 60 College Street, 606 LEPH, New Haven CT 06520
serap.aksoy@yale.edu,

²Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK,

³The Institute for Genomics Research, Rockville, MD, USA,

⁴RIKEN Genomic Sciences Center, Kanagawa, Japan,

⁵South African National Bioinformatics Institute (SANBI), Belville, South Africa

⁶Liverpool School of Tropical Medicine, Liverpool, L3 5QA UK

Keyword: Tsetse, *Glossina*, genome sequence, functional genomics

Teaser: Advancements in *Glossina* genomics have the potential to develop novel tools for African trypanosomiasis management. An international consortium has been formed to initiate efforts towards obtaining the full genome sequence of this important vector.

Abstract: Given the medical and agricultural significance of *Glossina*, knowledge on the genomic aspects of the vector and vector-pathogen interactions are a high priority. In preparation for a full genome sequence initiative, an extensive set of ESTs has been generated from tissue specific normalized libraries, BAC clones are being constructed and information on the genome structure and size from different species has been obtained. An international consortium is now in place to further efforts to lead to a full genome project.

Both male and female tsetse flies (Diptera:Glossinidae), are the cyclical vectors of the trypanosomes which cause African sleeping sickness in humans (HAT) and nagana in animals (AAT). It is conservatively estimated by the World Health Organization (WHO) that there are currently 300,000 – 500,000 cases of HAT with 60 million people at risk in 37 countries covering ~40% of Africa (11M km²) [1]. After a devastating epidemic in the early 20th century when a million people died of HAT, the disease almost disappeared from Africa by the 1960's. But we are now in the midst of another epidemic with increasing number of new infections and mortality (55,000 deaths in 1993; 66,000 in 1999) and a disease burden of 2.05M DALY (disability adjusted life years) [1-4]. The breakdown of surveillance, allied to displacement of populations by war and natural disaster, are contributory factors to this new epidemic. Given that HAT affects hard-to-reach rural populations, and that these war-torn areas lack active surveillance, the disease prevalence numbers are undoubtedly a gross underestimation. The considered view is that the situation may worsen [5,6]. In addition to the public health impact of HAT, it has been estimated that AAT limits the availability of meat and milk products in large regions of Africa. It also excludes effective cattle rearing from ten million square kilometers of Africa [7] with wide implications for land use; i.e., constraints on mixed agriculture and lack of animal labor for ploughing [8]. The programme on African animal trypanosomiasis (PAAT) estimate that AAT causes approximately 3 million cattle deaths per year and farmers are required to administer approximately 35 million doses of costly trypanocidal drugs many of which fail because of resistance in parasites developed for these chemicals [9-11]. Economic losses in cattle production are estimated at US\$ 1-1.2 billion annually and total agricultural losses are estimated at US\$ 4.75 billion per year (www.fao.org/ag/againfo/programmes/en/paat/home.html).

Transmission of trypanosomiasis requires four interacting organisms: the human host, the insect vector, the pathogenic parasite and the domestic and wild animal reservoirs. HAT is caused by the protozoan *Trypanosoma brucei rhodesiense* in East Africa and *T. b. gambiense* in West and Central Africa. The related trypanosomatids, *T. vivax* and *T. congolense*, are regarded as major pathogens of cattle and other ruminants, while *T. simiae* causes high mortality in domestic pigs. *T. brucei* affects all livestock, with particularly severe effects in equines and dogs. Members of both the *morsitans* and *palpalis* complex are efficient vectors for HAT and AAT. The recent epidemics caused by *T. b. gambiense* in Central Africa are transmitted by the *palpalis* group tsetse species *Glossina fuscipes*, which is also an important vector for *T. b. rhodesiense* in East Africa. The savannah species of tsetse in the *morsitans* complex, such as *Glossina morsitans morsitans* are also involved in the transmission of *T. b. rhodesiense* while being potent vectors for animal diseases as well.

The African trypanosome has been most studied in efforts to develop a mammalian vaccine for disease control (Box I). Recently, through a collaborative international effort, the two largest chromosomes of *T. b. brucei* have been sequenced and others are in the final stages of assembly (<http://www.genedb.org>) [12]. Realizing the importance and previous neglect of African trypanosome species other than *T. brucei*, whole genome shotgun sequencing projects for *T. congolense* and *T. vivax* have been recently initiated at the Wellcome Trust Sanger Institute (WTSI). *T. vivax* and *T. congolense* have so far been sequenced to six-fold and three-fold coverage, respectively, with the coverage of *T. congolense* expected to rise to five-fold shortly. While both the genomics and functional genomics aspects of trypanosomes are being

extensively explored and the human host has been fully sequenced, information on the genomics aspects of tsetse biology has been sparse.

The availability of genomics information can lead to the development of new control strategies aimed directly at the fly or at its parasite transmission ability (vector competence). The data from ongoing studies indicate that knowledge on tsetse-trypanosome interactions during the establishment of infections in the fly is central for the development of such strategies [13,14]. Through a symbiont-based transformation system (paratransgenesis), it has been possible to express foreign gene products in tsetse midgut [15-17]. To harness this system, it is now important to characterize trypanosome inhibitory products to express via paratransgenesis in the midgut milieu. Among such products are the tsetse immune system components, which will be elucidated through the genomics project [18,19]. The interactions of parasites with tsetse salivary gland tissue/saliva still remain a black box and the salivary gland expressed sequence tags (ESTs) project presently completed should provide a major resource for this important research area. The recently developed tool of gene silencing via dsRNA interference has been found to function successfully in tsetse (S. Aksoy and M.J. Lehane, unpublished). Hence, results beginning to be obtained from genomics studies could lead to functional analysis to better understand tsetse biology and parasite transmission to interfere with disease [20]. Another near-term benefit of genomics will be its impact on our knowledge of vector population biology. This information has the immediate potential to improve the efficacy and implementation of the current control programs on the ground [21]. Genomics information would also enable the development of new vector control initiatives including potential targeted insecticide development, host-seeking studies leading to improved trap and target design. Any approach relying on paratransgenesis and population replacement will also require a good understanding of vector populations and dynamics.

Finally, despite advances in the field of vector genomics, the small size of the tsetse research community remains a key obstacle to advancing research. Unfortunately, given the lack of funds, especially in African laboratories, many centers have disengaged their tsetse research programs and there are few laboratories in the developed world that currently maintain colonies and engage in research on tsetse. During the last decade, the World Health Organization has invested heavily in bio-informatics training courses in Africa, Latin America and Asia. In a recently announced initiative, the South African government has pledged funds over ten years to create a national bioinformatics network to support novel genome annotations as related to health in South Africa [22]. The involvement of African scientists and governments in the *Glossina* genome project would facilitate the application of various disease control tools anticipated to be developed from this new knowledge. A larger scientific community will help generate a research resource development and access facility in addition to promoting training and capacity building in disease endemic countries. Finally, comparative analysis of *Glossina* genes with their homologues in other diptera, *Drosophila*, *An. gambiae* and *Ae. aegypti*, would be welcomed by the larger vector biology community as they will shed light on the evolutionary processes that are conserved and play a role in invertebrate immunity in general [23].

Outline of proposed tsetse genomics activities

The genomics of tsetse can be planned in three phases, as described below.

Phase I consists of obtaining information on the genome size of *Glossina* species, cloning and sequencing of an extensive set of ESTs as part of a gene discovery project, construction of BAC (bacterial artificial chromosome) libraries and the preliminary sequencing of BAC-ends, and

sequencing of the tsetse symbiotic bacteria. Phase II would build on the preliminary information obtained above and aim to achieve partial three-fold shotgun coverage of the genome sequence, scaffolded together using BAC-end sequences. Furthermore, sequencing a small number of BAC clones would also allow us to better understand the complexity of this genome with respect to the organization of coding and non-coding regions and distribution of the repetitive elements. In addition, functional analysis of the tsetse transcriptome and its relation with its symbiotic flora will mediate a better understanding of the parasite transmission mechanism(s). Finally, Phase III would be full genome sequencing. Status of the ongoing work is presented below.

Analysis of EST libraries. Considerable progress has been made in generating ESTs and full-length gene sequences from *G. m. morsitans* (Table 1). Three tissue specific libraries, midgut, fat body and salivary gland, prepared following normalization of mRNA to reduce redundancies have produced over 67,000 tags. Generation of large-scale EST data is important as they represent an opportunity to identify genes expressed collectively among various developmental stages and in adults, and thus are generally reflective of the entire transcriptome. In addition, analysis of transcripts from specific tissues in response to trypanosome challenge and infections will provide the opportunity to identify important genes that are expressed in response to infection, hence can help understand the basis of vector competence. These libraries now represent a valuable immediate community resource and will permit subsequent full-length sequencing and can also be used for furthering functional genomics studies via microarray analysis. They will also be useful from a comparative perspective to understand the functional genomics of different vectors. Lastly, they will be important for training gene-finding software and subsequent annotation of the full genome.

Genome size of Glossina species. Members of both the *palpalis* and *morsitans* species complexes are important vectors of trypanosomes and there are good justifications for investigating the genome sequence from both species complexes. The laboratories of Spencer Johnson (Texas A&M) and Biemont Christian (Université Lyon1, France) have independently investigated the genome size of several *Glossina* species using FaxCalibur flow cytometer (shown in Table 2). Their results indicate that the genome of different species varies from 500-600 Mb in size, approximately 1.5 times the size of *Drosophila virilus* genome. Interestingly, William Black (U. Fort Collins, Colorado) has used reassociation kinetics analysis to determine the genome size of *G. p. palpalis*, which predicted a much larger size estimate of over about 7000 Mb. This analysis showed that 35% of the *G. p. palpalis* genome corresponds to foldback DNA, indicating the presence of a large heterochromatic region. It is possible that the large load of the bacterial symbionts associated with tsetse may have been a confounding factor in the size variation observed between the two techniques used.

Genomic DNA libraries. BAC libraries for *G. m. morsitans* are currently being constructed, funded through NIH/NHGRI (<http://www.genome.gov/10001852>) and the Wellcome Trust Sanger Institute (WTSI). The desired average insert size of the libraries will be around 120-140 kb, with an overall genome coverage of approximately ten-fold. Funds have also been made available for the sequence analysis of 60,000 BAC-ends and for complete sequence of several BAC clones through RIKEN Genomics Sciences Center (RIKEN GSC) and WTSI. These paired BAC-end sequence data are important as sequence-tagged-connectors to assist in linking contiguous sequences together, into longer chains or scaffolds. These data also will provide

valuable preliminary information on genome structure including, for example, repetitive element type, frequency and variability as well as putative coding sequence frequency.

Genomics of tsetse symbiotic bacteria. Three microbial organisms closely interact and influence tsetse physiology. These symbiont genomes are of interest since in the absence of their gut flora, tsetse flies are severely impaired in important physiological functions such as longevity and reproduction [24]. The bacteria are also of interest as they have been implicated in modifying the vector competence of their host [25]. Two of the symbionts are enteric, genus *Sodalis glossinidiae* and *Wigglesworthia glossinidia* [16]. The *Wigglesworthia* [26] and *Sodalis* genomes (near completion) have been sequenced at RIKEN GSC. Work with microarrays is in progress in order to investigate the functional genomics aspects of the tsetse-symbiont interactions (Aksoy laboratory, Yale University). The third symbiont *Wolbachia* can confer mating incompatibilities to various insects it infects, which results in the spread of the infected insect phenotypes in the field. *Wolbachia* has also invaded many natural tsetse populations and such mating incompatibilities it might confer have the potential to drive engineered parasite refractory tsetse into natural populations as an alternative strategy for disease control [16,27,28].

Phase II resources needed. Additional resources will enable the genomics initiative, and allow for full exploitation of the existing genomics information. These include the construction of a full-length cDNA library to expand the gene discovery studies. With an extensive set of BAC clones now being partially sequenced, a physical map of tsetse chromosomes is needed to facilitate the scaffolding. Physical mapping in tsetse can be based on *in situ* hybridization to metaphase chromosomes, as its genome organization (likely due to the repetitive nature) is not conducive to suitable polytene chromosome preparations. Alternatively a BAC restriction mapping approach may allow for additional genomic information. These resources would facilitate the development of functional genomics studies to ensure that genomics data can result in applications that lead to disease reduction. In particular, the construction of microarrays and their application for gene expression analysis would immediately benefit the host-pathogen studies aimed to interfere with trypanosome transmission in tsetse.

Phase III studies. Finally, the community would be ready to undertake the genome sequence project using the whole-genome shotgun approach with an eight to ten fold coverage that will lead to a complete annotated genome of *Glossina*. We would hope that random paired-end sequences providing eight-fold redundant coverage would produce an assembly of over 90% of the euchromatin of the tsetse genome. The preliminary information gathered from *Phase II* will ultimately influence the final strategy used to obtain a full sequence. For instance, the balance of benefits between clone-by-clone and whole genome shotgun sequencing are influenced by clone library representation and genome polymorphisms, respectively. Physical and FISH (fluorescence *in situ* hybridization) mapping of BACs and their fingerprints will allow large scaffolds to be mapped to chromosomes and will provide a framework for any future gap closure efforts. All information can then be deposited to a general vector database to benefit the larger vector biology groups. One such database, VectorBase has recently been initiated and the inclusion of *Glossina* information on this network would be desirable (Frank H. Collins, University of Notre Dame, personal communication).

Community interest in a *Glossina* genome project. In an effort to review the status of genomics resources in *Glossina* and further its development, a small meeting was held in January 2004 in Geneva, under the auspices of the World Health Organization's Tropical Disease Research (TDR) programme. This meeting brought together scientists with molecular interests from about a dozen sleeping-sickness labs and genome centers in an effort to promote the genomics activities [22]. A mail server has been established at the South African National Bioinformatics Institute to ensure efficient communication and distribution of information to interested scientists (www.glossina.sanbi.ac.za). The group, International *Glossina* Genome Initiative (IGGI) will hold its second meeting in February 2005 to review progress and prospects towards the full genome sequence.

Box I. Trypanosomiasis control

The African trypanosome has been most studied in efforts to develop a vaccine for immunization of humans and cattle. Unfortunately, antigenic variation in the trypanosomes while in the mammalian host has hampered efforts for vaccine development with no effective products forthcoming for disease control in the foreseeable future. Current disease management primarily depends on active surveillance and treatment of patients and some vector control initiatives. The drug treatment of HAT is in a parlous state [29-31]. It relies on old, often dangerous drugs, and high levels of parasite resistance is emerging as a major problem. Recently synthesis of suramin and eflornithine used for treatment in the early stages of disease was about to be terminated by their manufacturers and was only saved at the last minute by an International outcry [32]. The successful experimental results obtained with DB333 derivatives for *T. b gambiense* stage-2 disease is very welcome news. In contrast, no new drugs are in the pipeline for the end-stage disease. Melarsoprol, an arsenical drug, has been used as first-line treatment for late-stage HAT for several decades. The drug is dangerous in its own right producing a commonly fatal reactive encephalopathy in about 5% of patients. Furthermore, the reported parasite resistance to melarsoprol is alarming with at least 20% of patients not responding to Melarsoprol in this epidemic [11,33]. It is hoped that the parasite genome sequence information now available will provide the impetus and opportunities for the identification of unique targets for which effective drugs can be developed.

In the absence of vaccines and effective and affordable drugs, disease control, via the control of its insect vector has been found to be highly effective. It is likely that tsetse control will remain one of the most effective overall approaches to the control of African trypanosomiasis. Current vector control efforts center largely on trapping, or killing the tsetse with insecticides [34,35]. While these efforts are effective, they have been difficult to sustain at the local community level [36]. The reduction of populations by sterile insect technique (SIT), which has an area-wide impact has been found to be highly effective [37]. Findings from genomics studies have the potential to significantly improve upon the existing control strategies. Knowledge of genes related to host-parasite interactions are vital for the development of genetically engineered lines that are unable to transmit trypanosomes, which can be immediately used in the ongoing SIT release programs. This application of refractory strains in SIT would reduce the cost of the projects and also increase the efficacy of their application in HAT endemic areas [15]. Undoubtedly genes related to olfaction can result in enhancement of trapping

technologies, especially in the case of the human disease transmitting *palpalis* group flies for which efficient trapping systems do not exist.

Table 1. EST collections either currently available, in progress, or in planning stages.

Tissue source of library	# ESTs	Known genes	Investigator/status
Normalized midgut from naïve and trypanosome infected <i>G.m morsitans</i>	21,427	4,035	www.sanger.ac.uk/Projects/G_morsitans
Normalized salivary gland from <i>G. m. morsitans</i>	27,426	5,895	www.sanger.ac.uk/Projects/G_morsitans
Normalized fatbody from naïve and immune challenged <i>G. m. morsitans</i>	20,257	6,372	www.tigr.org in final annotation stage
Developmental stages of <i>G. m. morsitans</i>	5,000		S. Aksoy/ in progress
Antenna - <i>G. m. morsitans</i> ; <i>G. p. palpalis</i>	3,000		M. J. Lehane/in progress
Adult naïve – full length <i>G. m. morsitans</i>	10,000		S. Aksoy/RIKEN GCG/planned
Fatbody and gut - <i>G. p. palpalis</i>	10,000		Genoscope/in progress

Table 2. Genome size estimates of different *Glossina* species

<i>Glossina</i> species analyzed	Haploid genome size (pg or Gb)	Ratio <i>Glossina/D. virilis</i> *
<i>Glossina morsitans morsitans</i>		
Male	0.579 (0.590 [^])	1.546
Female	0.613 (0.596 [^])	1.634
<i>Glossina pallidipes</i>		
Male	0.509	1.356
Female	0.533	1.422
<i>Glossina palpalis palpalis</i>		
Male	0.482	1.285
Female	0.479	1.278
<i>Glossina fuscipes</i>		
Male	0.534	1.523
Female	0.524	1.398

*The haploid genome size of *D. virilis* has been estimated as 0.34–0.38 pg

[^]Values independently determined by Dr. Spencer Johnston, Texas A&M.

All values determined using FaxCalibur flow cytometer by Biemont Christian, Christiane Nardon and Michèle Weiss, at Université Lyon1, France

References:

- 1 Committee, T.S. (2001) *Scientific working group on African trypanosomiasis (sleeping sickness)* WHO/Tropical Disease Research Unit
- 2 Moore, A. et al. (1999) Resurgence of sleeping sickness in Tambura County, Sudan. *Am J Trop Med Hyg* 61 (2), 315-318
- 3 Ekwanzala, M. et al. (1996) In the heart of darkness:sleeping sickness in Zaire. *Lancet* 348, 1427-1430
- 4 van Hove, D. (1996) Sleeping sickness in Zaire. *Lancet* 349, 438
- 5 Barrett, M. (1999) The fall and rise of sleeping sickness. *Lancet* 353 (1113-1114)
- 6 Smith, D.H. et al. (1998) Human African trypanosomiasis: an emerging public health crisis. *British Medical Bulletin* 54, 341-355
- 7 Steelman, C.D. (1976) Effects of external and internal arthropod parasites on domestic livestock production. *Annual Review of Entomology* 21 (4), 155-178
- 8 Jordan, A.M. (1986) *Trypanosomiasis control and African rural development*, Longman
- 9 Afewerk, Y. et al. (2000) Multiple-drug resistant *Trypanosoma congolense* populations in village cattle of Metekel district, north-west Ethiopia. *Acta Tropica* 76 (3), 231-238
- 10 Bacchi, C. (1993) Resistance to clinical drugs in African trypanosomes. *Parasitology Today* 9 (5), 138-145
- 11 Geerts, S. et al. (2001) African bovine trypanosomiasis: the problem of drug resistance. *Trends Parasitol* 17 (1), 25-28.
- 12 El-Sayed, N.M. et al. (2003) The sequence and analysis of *Trypanosoma brucei* chromosome II. *Nucleic Acids Res* 31 (16), 4856-4863
- 13 Aksoy, S. et al. (2002) What can we hope to gain for trypanosomiasis control from molecular studies on tsetse biology ? *Kinetoplastid Biol Dis* 1 (1), 4
- 14 Aksoy, S. (2003) Control of tsetse flies and trypanosomes using molecular genetics. *Vet Parasitol* 115 (2), 125-145
- 15 Aksoy, S. et al. (2001) Prospects for control of African trypanosomiasis by tsetse vector manipulation. *Trends in Parasitology* 17 (1), 29-35
- 16 Rio, R.V. et al. (2004) Strategies of the home-team: symbioses exploited for vector-borne disease control. *Trends Microbiol* 12 (7), 325-336
- 17 Cheng, Q. and Aksoy, S. (1999) Tissue tropism, transmission and expression of foreign genes in vivo in midgut symbionts of tsetse flies. *Insect Molecular Biology* 8 (1), 125-132
- 18 Hao, Z. et al. (2001) Tsetse immune responses and trypanosome transmission: implications for the development of tsetse-based strategies to reduce trypanosomiasis. *Proc Natl Acad Sci USA* 98 (22), 12648-12653
- 19 Hao, Z. et al. (2003) Proventriculus (cardia) plays a crucial role in immunity in tsetse fly (Diptera: Glossinidae). *Insect Biochem Mol Biol* 33 (11), 1155-1164
- 20 Lehane, M.J. et al. (2003) Adult midgut expressed sequence tags from the tsetse fly *Glossina morsitans morsitans* and expression analysis of putative immune response genes. *Genome Biol* 4 (10), R63
- 21 Krafsur, E.S. (2003) Tsetse fly population genetics: an indirect approach to dispersal. *Trends Parasitol* 19 (4), 162-166
- 22 Butler, D. (2004) African labs win major role in tsetse-fly genome project. *Nature* 427 (6973), 384

- 23 Lehane, M.J. et al. (2004) Immune responses and parasite transmission in blood-feeding insects. *Trends Parasitol* 20 (9), 433-439
- 24 Nogge, G. (1976) Sterility in tsetse flies (*Glossina morsitans* Westwood) caused by loss of symbionts. *Experientia* 32 (8), 995-996
- 25 Welburn, S.C. and Maudlin, I. (1999) Tsetse-trypanosome interactions: Rites of passage. *Parasitology Today* 15 (10), 399-403
- 26 Akman, L. et al. (2002) Genome sequence of the endocellular obligate symbiont of tsetse, *Wigglesworthia glossinidia*. *Nature Genetics* 32 (2), 402-407
- 27 O'Neill, S.L. et al. (1993) Phylogenetically Distant Symbiotic Microorganisms Reside in *Glossina* Midgut and Ovary Tissues. *Medical and Veterinary Entomology* 7 (4), 377-383
- 28 Cheng, Q. et al. (2000) Tissue distribution and prevalence of *Wolbachia* infections in tsetse flies, *Glossina* spp. *Medical and Veterinary Entomology* 14 (1), 44-50
- 29 Butler, D. (2003) Tropical diseases: raiding the medicine cabinet. *Nature* 424 (6944), 10-11
- 30 Nok, A.J. (2003) Arsenicals (melarsoprol), pentamidine and suramin in the treatment of human African trypanosomiasis. *Parasitol Res* 90 (1), 71-79
- 31 Etchegorry, M.G. et al. (2001) Availability and affordability of treatment for Human African Trypanosomiasis. *Trop Med Int Health* 6 (11), 957-959
- 32 McNeil, D. (2000) Drug Companies and Third World: A case study in neglect. In *New York Times*, pp. 1
- 33 Anene, B.M. et al. (2001) Drug resistance in pathogenic African trypanosomes: what hopes for the future? *Vet Parasitol* 96 (2), 83-100.
- 34 Joja, L.L. and Okoli, U.A. (2001) Trapping the vector: community action to curb sleeping sickness in southern Sudan. *Am J Public Health* 91 (10), 1583-1585.
- 35 Lancien, J. (1991) Campaign against sleeping sickness in South-West Uganda by trapping tsetse flies. *Ann Soc Belg Med Trop* 71 Suppl 1, 35-47
- 36 Hursey, B.S. (2001) The programme against African trypanosomiasis: aims, objectives and achievements. *Trends Parasitol* 17 (1), 2-3.
- 37 Vreysen, M.J. et al. (2000) *Glossina austeni* (Diptera: Glossinidae) eradicated on the Island of Unguja, Zanzibar, using the sterile insect technique. *J. Econ. Entomology* 93, 123-135