The transcriptome profile of *Glossina morsitans morsitans*: a vector for sleeping sickness

Mark Wamalwa[1], IGGI[2] & Alan Christoffels[1] *

[1]Faculty of Natural Sciences, University of the Western Cape,
South African National Bioinformatics Institute (SANBI), Private Bag X17,
Bellville 7535, South Africa.
[2]The International Glossina Genome Initiative (IGGI).
*To whom correspondance should be addressed
e-mail: alan@sanbi.ac.za, markw@sanbi.ac.za

Glossina morsitans morsitans (Diptera: *Glossinidae*) is a haematophagous insect vector responsible for transmitting African trypanosomes, the causative agent of human African trypanosomiasis (HAT) and 'nagana' in animals. Despite being one of the most important organisms for transmitting 'sleeping-sickness', the lack of available information about its genetic makeup limits research with this vector. Over the past 5 years, the International Glossina Genome Initiative has coordinated the generation of 125,000 expressed sequence tags (ESTs) from eleven cDNA libraries. We analyzed the transcriptome of *Glossina* (tsetse) based on homology to gene products of *Aedes aegypti*, *Anopheles gambiae*, *Apis mellifera*, *Tribolium castaneum*, *Drosophila melanogaster* and *Ixodes Scapularis*. Using the Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG), we present a functional classification of the tsetse-fly transcriptome. Two computational strategies are emphasized (1) Markov Chain clustering of orthologs and (2) Statistical gene-set enrichment of GO terms.

Data was sourced from Ensembl release 49 and VectorBase. An automated comparative analysis pipeline was applied to identify coding potential of ESTs. GO functional profiling was applied and significant GO-terms (p-value < 0.05) were identified using Fisher's exact test preceded by Bonferroni correction. Putative secretory ESTs were predicted using SignalP version 3 while KEGG was queried to cluster genes into metabolic pathways.

More than 20% (4,680) of the transcriptome is not represented in sequenced insect genomes while 8% (1,338) constitute the conserved core. About 19% (7,205) of the transcripts have homologues to proteins of known structure. Functional classification identified novel tsetse-specific features such as enrichment of iron binding proteins, proteases (anticoagulant) and immunity-associated transcripts. These findings underscore the importance of the *G. morsitans* transcriptome for comparative analysis of the human host, the insect vector and the pathogenic parasite.

Keywords: expressed sequence tags; database similarity searches; functional annotation; gene structure