Tsetse questions for the Press Release

Describe, in general terms, what has been accomplished.

The genetic code for a representative species of the tsetse fly (*Glossina morsitans*) has been sequenced and annotated. Tsetse flies function as the insect vectors of African trypanosomes to humans and animals in sub-Saharan Africa. Tsetse are also known for their unique biology in that they feed exclusively on vertebrate blood, give birth to live young, provide nutrition to their young by lactation and have formed complex relationships with 3 different symbiotic bacteria. The tsetse genomic sequence was produced and annotated during a ten year collaborative effort by an international scientific collaboration. A primary goal of the analysis of the genome focused upon identifying and annotating the genes within the genome sequence which code for the proteins that make up the tsetse fly. The proteins produced from the genome of an organism are involved in every aspect of its structure and function from the earliest stages of development to its mature form. They are basically the parts list that an organism is made from. Protein coding genes were identified and mapped to the tsetse genome by using computer programs that "read" the sequence and compare the massive amounts of sequence data from the genome with that from other annotated organisms to predict gene structure and function. These predictions were then added to the publicly available tsetse genomic database found at www.vectorbase.com.

The automated annotations were then closely examined and manually curated by a group of ~140 scientists within the tsetse and insect disease vector biology communities who specialize in a wide array of topics important to understanding the biology of the tsetse fly. These topics include olfaction (smell), gustation (taste), vision, reproduction, digestion, blood feeding, immunity, metabolism, stress response, symbiotic relationships between tsetse and its associated microbial entities, and the hormonal regulation of genes and physiological functions. The analyses performed by these research groups were used to update the automated predictions and add information to the gene predictions in the database. The process of annotation is ongoing and information will be continually added to this resource over time.

The resulting genome paper summarizes a handful of the findings produced by these analyses within the context of the most significant aspects of tsetse's unique biology and in its role as a disease vector. More specific and in depth analyses regarding various aspects of tsetse's biology are being published in coordination with the genome paper as a collection of 10 papers within the PLoS family of journals.

How will this information be useful? How will it be used? What is its potential?

Detection and treatment of trypanosomiasis is expensive, difficult and dangerous. Disease prevention by vaccine development does not appear feasible due to the ability of trypanosomes to evade the mammalian immune system. Currently, African trypanosomiasis is controlled primarily by vector (tsetse fly) control through techniques such as trapping, pesticide treatments and sterile male release strategies. The information contained within the genome provides a foundational resource which will be invaluable to the tsetse and insect vector biology communities. Access to this information will accelerate research on tsetse's basic biology. The outcomes of this research can then be applied to improving current tsetse control methods and the development of new strategies with an emphasis upon improving effectiveness and cost reduction.

Beyond disease control, the genome is an important resource for evolutionary biology. Tsetse flies are unique from the majority of other insects in multiple aspects of their biology. They provide examples of dramatic physiological adaptations which include

- Independent evolution of blood feeding (different from mosquitoes)
- Development of an essential relationship with a bacterial symbiont (*Wigglesworthia*) required for multiple aspects of their biology (immunity, metabolism and reproduction)
- Use of the amino acid proline instead of sugars for energy transport and metabolism
- Evolution of intrauterine development of offspring and live birth
- Independent evolution of lactation for nourishment of intrauterine offspring

The evolution of these amazing adaptations can now be examined on a genomic level relative to other related insects (such as the fruit fly *Drosophila*) for which genomic information is also available. Insights gained from such comparisons allow us to understand how such dramatic changes develop at the genetic level in related organisms.

It also provides an opportunity to study convergent evolution in which two distantly related organisms develop similar biological adaptations. An example of this is tsetse's evolution of lactation. Lactation is a defining aspect of mammalian biology. The independent evolution of this physiology in tsetse allows the comparison of the commonalities and differences between the two systems. Surprisingly, analyses of proteins in tsetse milk reveal significant similarities to those found within mammalian milk!

Describe the biggest challenge faced during this process and how it was overcome.

- Since the project was initiated on a shoe-string, it was a challenge to keep the consortium together and to raise funds for the project and associated activities. The older methods were much more expensive than the new approaches used now.

- On the final analysis, the challenge was to assemble the information from many groups and distilling it into a coherent paper, but I don't think that that was the biggest challenge this process faced. It was a lot of work, but I think the biggest hurdles to overcome arose from funding, acquiring the required materials and the actual sequencing.

Outline the logistics behind this accomplishment (e.g., time, # of people, expense, etc.). How was it funded?

- Project began in 2003 when the International Glossina Genome Initiative was established with seed funds from the WHO
- ~146 scientists have been associated with the project – at least half are from African institutions
- The cost of the larger project, including training efforts of IGGI, genome sequence at Sanger and the functional genomic studies performed in various labs building on the genome data exceeded 10 million US $
- The project was funded through multiple resources, the consortium was funded by the WHO/TDR, BAC and EST libraries were funded by NIH and Wellcome Trust, the bulk of the genome sequencing effort was funded by Wellcome Trust with contributions from RIKEN and Genoscope

Why does the process take so long?

- Removal of symbiont genomic material was essential for the project thus obtaining sufficient amount of material from one female line was difficult given that tsetse give rise to few progeny
- Assembly of the genome was difficult as genomic material originated from multiple genetic lines due to restrictions relating to the small amount of DNA obtained per fly and the small number of offspring per fly
- Technologies utilized early in the project required more raw materials and yielded less/lower quality sequences
- The older technologies were more costly and labor intensive than the new methods recently developed
- Newer technologies facilitated completion of the genome
- Getting a community of experts for the final annotation step was challenging and coordination of a project with many people required extensive communication and organization
- The amount of data produced by a genome project is huge and requires a lot of time for adequate annotation by both computers and humans

How many other species have been similarly sequenced? How much data does the tsetse sequencing translate into?

- Currently, many insect genomes are becoming available including those from multiple mosquito species, fruit fly species

- Tsetse genome is one of the first from this subcategory of fly species (Suborder Brachycera) with the exception of multiple fruit fly genomes.
- 5 more tsetse genome projects are in progress as well as the related house fly (Musca domestica) and the stable fly (Stomoxys Calcitrans).
- The size of the tsetse fly genome is ~366 Megabase pairs (~366 million letters of code)
- Roughly twice the size of the fruit fly (*Drosophila*) genome and 10% of the size of the human genome.