A tale of three genomes: the kinetoplastids have arrived

Jessica C. Kissinger
Department of Genetics and Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA 30606, USA

July 2005 marked a milestone in kinetoplastid biology research. A tour de force effort led by the Tri-Trypanosomatidae ‘Tritryp’ genome consortium yielded the publication of three prominent kinetoplastid parasite genome sequences: Trypanosoma brucei, Trypanosoma cruzi and Leishmania major. The individual and combined comparative analyses of these three genome sequences, combined with proteomic analyses, have yielded insights into topics ranging from genome evolution and horizontal gene transfer to potential new therapeutic and vaccine targets.

The Tritryp genome sequences and the Trypanosoma cruzi proteome
The publication of three draft kinetoplastid genome sequences [1–3] and a proteomic expression study [4] for kinetoplastid parasites is a significant research advance in the quest for therapeutic and vaccine targets for the causative agents of leishmaniasis, African sleeping sickness and Chagas disease. The generation of these genome sequences and their subsequent analysis by the Tritryp genome consortium represent an international scientific collaboration worth taking note of. The three genome projects originated in different countries, with different sequencing strategies, and focused on organisms that varied in their particular biological and technical challenges. Leishmania major and Trypanosoma brucei, the first genome projects to be started, were sequenced in a chromosome-by-chromosome approach. The T. cruzi project began with a ‘map-as-you-go’ strategy because of its complex karyotype, but had to switch to a whole-genome shotgun (WGS) approach after the highly repetitive nature of the genome was discovered. These genomes were also sequenced as diploids, and the allelic variation present in the hybrid T. cruzi CL Brener strain challenged and hampered assembly efforts. Yet, given the added value of the comparative approach, these projects came together under the Tritryp consortium umbrella and benefited from the strengths each provided. In addition, proteomic expression analyses of multiple life-cycle stages in T. cruzi have provided a comprehensive view of protein expression in a kinetoplastid organism and have also
served as an independent validation of genome sequence annotation.

The group of papers reporting on these projects – the genome sequences of *T. brucei* [1], *T. cruzi* [2] and *L. major* [3], the proteome of *T. cruzi* [4], and a comparative genomic analysis of trypanosomatid parasitic protozoa [5] – should be considered together (including the genes in the supplementary material that accompanies them). The data analysis among the projects is so integrated and comparative that each paper focuses on a few distinct biological areas but does so in the context of all three organisms, with the exception of the *T. cruzi* proteome, which was a distinct research effort. These papers represent our first detailed look into the kinetoplastid genome.

**Comparative genomic insights – the kinetoplastid core**

Comparative analyses have revealed that all three kinetoplastid genomes share 6158 ortholog clusters of protein-encoding genes and many functional protein domains (Figure 1). Moreover, these genes exist in large syntenic blocks that contain 80% of the *T. brucei* and 93% of the *L. major* genes. This is a remarkable degree of conservation for organisms that diverged an estimated 200–500 million years ago and differ in their vector, mechanism of immune evasion, target tissue and pathogenesis.

The high level of syntenic conservation among these three genomes has had two extremely beneficial outcomes: (i) it has greatly facilitated annotation efforts for each of the genomes; and (ii) it has made the gene content and genomic structure differences between these organisms easier to dissect. Mobile DNA elements, structural RNAs and many species-specific genes, including known surface antigens, appear to reside in the break points between syntenic regions [5].

**Genomics, parasitism and pathogenesis**

Insights into the origin of parasitism in the trypanosomatids [6], the evolution of mechanisms of immune evasion and a determination of pathogenesis factors can be gleaned from examination of the three kinetoplastid genome sequences. Although *L. major*, *T. brucei* and *T. cruzi* are all parasites, they differ significantly in their interactions with hosts. For instance, *T. brucei* and *T. cruzi* have energy-generating pathways that rely on either proline or histidine, respectively, and these pathways have made them exquisitely adapted to niches in different orders of insects as vector hosts. The apparent acquisition of nearly 50 bacterial enzymes by some, if not all, of the Trytrop species [1] has provided the parasites with some novel capabilities and has provided us with potential drug targets. As an example, consider the attachment of mucins in *T. cruzi* to glycosylphosphatidylinositol (GPI) anchors through aminoethylphosphonate. Only *T. cruzi* contains this complete pathway, and all components are absent from humans, thus enzymes in this pathway could serve as potential drug targets. In addition, this pathway is particularly attractive because the last step, aminoethylphosphonate transaminase, is a bacterial enzyme acquired by horizontal gene transfer in both *L. major* and *T. cruzi* [1].

An examination of the genome sequences for insights into mechanisms of immune evasion has revealed that ~20% of the *T. brucei* genome is species-specific (Figure 1) and dedicated to antigenic variation. Most variant surface glycoproteins (VSGs) are associated with upstream repeat elements, exist in subtelomeric arrays and are defective. Only 57 out of 806 analyzed VSGs are fully functional [1]. In *T. cruzi*, the number of species-specific genes also appears to be greatly expanded with respect to the other species: 3736 versus 1392 (*T. brucei*) and 910 genes (*L. major*) (Figure 1). A new surface molecule family, the mucin-associated surface proteins (MASPS), contains >1300 copies. In addition, the genome encodes 1430 trans-sialidases, 863 mucins and 565 gp63 surface proteases (not counting pseudogenes) [2]. As is the case with VSG genes, many of the genes encoding *T. cruzi* surface molecules are located in subtelomeric regions or in regions associated with synteny breaks between the species. The *L. major* genome also encodes a variety (although much smaller number) of surface molecules, including lipophosphoglycans, glycoinositolphospholipids, proteophosphoglycans and GPI-anchored proteins such as gp63 [3].

**Proteomic insights and vaccine prospects**

Several proteomic analyses have been performed on particular proteins, pathways and stages of kinetoplastid parasites over the past several years [7–10]; however, none has been of the scale reported for *T. cruzi*. Peptides mapping to 2784 distinct proteins belonging to 1168 annotated orthologous protein groups were detected in the proteomic analysis of four *T. cruzi* life-cycle stages [4]. In addition to providing our first detailed look into the patterns of protein expression in a kinetoplastid organism, the peptide hits served as an independent validation of the initial *T. cruzi* annotation. Only 79 new genes, new alleles or modifications to existing annotation were necessary as
a result of the proteomic data. In addition, evidence for the expression of ~1000 genes annotated as hypothetical (500 of which are conserved with at least one other species) was generated. Recently, a proteomic study in *T. brucei* procyclic stages provided similar evidence for the quality of the *T. brucei* annotation [11].

With respect to patterns of protein expression, the *T. cruzi* analysis revealed that ~30% of the detected genes were expressed in all stages examined and, importantly, that the parasite relies on distinct energy sources in different life-cycle stages, including histidine metabolism in insect stages. These and other stage-specific metabolic genes might serve as new potential therapeutic targets. Evidence for expression of the newly identified MASP gene family was found, primarily in trypomastigotes. Surface proteins expressed in abundance during amastigote and trypomastigote stages could serve as vaccine targets [4].

**More tools and genomes on the horizon**

Comparative genomics and post-genomic technologies hold great potential for further elucidating kinetoplastid biology [12]. A genetic map for *T. brucei* with 182 markers on all chromosomes linked to genome sequence [13] has recently been generated. This advance allows, for example, forward genetic approaches including positional cloning and the identification of drug resistance loci. Other functional approaches for *T. brucei* include attempts to identify mutants systematically through targeted gene inactivation using RNA interference (RNAi) (see TrypanoFAN, the *T. brucei* Functional Genomics Project: http://trypanofan.path.cam.ac.uk). Given that significant numbers of the predicted genes are hypothetical or conserved hypothetical [5], including many with evidence of protein expression in *T. cruzi* [4] and *T. brucei* [11], additional molecular, functional genomic and genetic approaches are needed to ascertain their function. The polycistronic nature of transcription in the Kinetoplastida, combined with the extensive use of post-transcriptional regulatory mechanisms, make microarray analyses challenging, but they have been performed [14].

Studies focused on elucidating parasite adaptations to host, immune evasion, drug resistance and evolution benefit greatly from comparative genomic approaches. The value of low coverage comparisons to other strains has been well demonstrated in other pathogens and parasites, particularly the malaria parasite *Plasmodium* [15–17]. This approach has also been adopted for the kinetoplastid parasites and a total of ten kinetoplastid genome projects of varying coverage for several *Leishmania* and *Trypanosoma* species are currently in progress or are planned (Table 1).

In this tale of three genomes, there are still many unknowns, but the stage is set. The efforts of the Tritryp genome consortium have established a bridge between the Tritryp research communities and have demonstrated an amazingly conserved genetic core that will open new avenues of research. Consider, for example, a conserved hypothetical protein found in all three genomes: its expression can be validated in *T. cruzi*, its phenotype assayed by RNAi in *T. brucei* and its relevance to host or tissue specificity gleaned from comparative genomics with more distant species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Status of data</th>
<th>Website</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Leishmania braziliensis</em></td>
<td>5X WGS; pre-finishing based on <em>L. major</em> backbone</td>
<td><a href="http://www.GeneDB.org">http://www.GeneDB.org</a></td>
<td>[18]</td>
</tr>
<tr>
<td><em>Leishmania infantum</em></td>
<td>5X WGS; pre-finishing based on <em>L. major</em> backbone</td>
<td><a href="http://www.sanger.ac.uk/Projects/L_braziliensis">http://www.sanger.ac.uk/Projects/L_braziliensis</a></td>
<td>[18]</td>
</tr>
<tr>
<td><em>Leishmania major</em></td>
<td>Completed. One telomere of chromosome (chr.) 8 appears as ‘missing’ but is likely identical to the other telomere</td>
<td><a href="http://www.GeneDB.org">http://www.GeneDB.org</a></td>
<td>[3,18]</td>
</tr>
<tr>
<td><em>Leishmania mexicana</em></td>
<td>DNA in preparation for sequencing</td>
<td><a href="http://www.sanger.ac.uk">http://www.sanger.ac.uk</a></td>
<td></td>
</tr>
<tr>
<td>Trypanosoma brucei brucei</td>
<td>Ongoing: chr. 2–8 finished, excluding telomeres and subtelomeres; chr. 1 and 10 finished; chr. 9 and 11 have a few gaps. A minichromosome library is in preparation and an intermediate chromosome library is planned. Telomere sequencing (TAR) cloning is planned. Proteomic data are available for procyclic stages</td>
<td><a href="http://www.GeneDB.org">http://www.GeneDB.org</a>; <a href="http://www.tigr.org/tdb/e2k1/tba1">http://www.tigr.org/tdb/e2k1/tba1</a> <a href="http://www.sanger.ac.uk/Projects/T_brucei">http://www.sanger.ac.uk/Projects/T_brucei</a> <a href="http://pantar.dcs.gla.ac.uk/RAPAD/Projects/index.html">http://pantar.dcs.gla.ac.uk/RAPAD/Projects/index.html</a></td>
<td>[1,11,19,18]</td>
</tr>
<tr>
<td><em>T. b. brucei</em> 427</td>
<td>TAR cloning</td>
<td><a href="http://www.sanger.ac.uk">http://www.sanger.ac.uk</a></td>
<td>[20]</td>
</tr>
<tr>
<td><em>T. b. gambiense</em></td>
<td>8X WGS; analysis and annotation available</td>
<td><a href="http://www.GeneDB.org">http://www.GeneDB.org</a></td>
<td>[18]</td>
</tr>
<tr>
<td>Trypanosoma congolense</td>
<td>5X WGS; automated annotation available</td>
<td><a href="http://www.GeneDB.org">http://www.GeneDB.org</a></td>
<td>[18]</td>
</tr>
<tr>
<td>Trypanosoma vivax</td>
<td>6X WGS; automated annotation available</td>
<td><a href="http://www.GeneDB.org">http://www.GeneDB.org</a></td>
<td>[18]</td>
</tr>
</tbody>
</table>

---

**References**

Letters

Malaria historians: analyze your old slides!

Carol Hopkins Sibley

Department of Genome Sciences, Box 357730, University of Washington, Seattle, Washington, 98195-7730, USA

Are you interested in the history of drug resistance in \textit{P. falciparum} or \textit{P. vivax}? To reconstruct that history, we all need to analyze old samples, and those are not easy to find.

Population genetic analyses of \textit{P. falciparum} have enabled us to begin to trace the historical and the geographical patterns of resistance to the widely used antimalarial drugs, chloroquine and sulfadoxine-pyrimethamine (SP). Initially, these studies examined only genes associated with resistance to the individual drugs: \textit{Pfcr}t and \textit{Pfmdr}1 for chloroquine and \textit{dhfr} and \textit{dhps} for SP. More recently, the completion of the genome sequence of \textit{P. falciparum} enabled extended regions of the genome to be examined using microarrays. For both chloroquine and SP, the data suggest that resistant genotypes have arisen relatively rarely and spread widely from the regions where they arose [1–3].

Currently, there are efforts in most endemic regions to measure the prevalence of alleles and extended haplotypes associated with resistance to these old drugs, and to identify genes that may serve as markers of resistance to the many new drugs and combinations being deployed. One goal of this work is to track the speed with which resistant genotypes are selected under drug pressure, or with which they decline in prevalence as the selective drug is replaced. This has been done in a few instances [4–6] and the data have been used in models to try to generalize the observed changes. However, there is considerable controversy about the role of other important factors in these trends. In particular, differences in drug usage, level of transmission and the mosquito vectors most active in the region are likely to have an important impact on the speed of the genetic changes, but a wide range of other ecological and demographic parameters are also important. It will require documentation of these trends from a wide variety of locations before any general conclusions can be drawn.

This is not just an academic exercise. Prediction of the useful therapeutic life of newly deployed drugs is crucial for sensible management of antimalarial drug use. At the moment, we have no way to tell whether we can extrapolate trends in allele prevalences observed in Thailand to make useful predictions in Tanzania or the Amazon basin. Historical samples could provide a baseline from which to measure the rates of change in allele prevalences after drugs are introduced in various areas. A systematic analysis of these datasets would allow us to begin to determine whether we can, in fact, extract general lessons from the different locations. National Malaria Control Programs, organizations such as the Global Fund to Fight AIDS, TB and Malaria, or the World Bank, that support drug purchases by endemic countries, and companies or organizations such as the Medicines for Malaria Venture that invest in antimalarial drug development all have a stake in the validity of these projections.

What do we need to begin to define these genetic patterns? Determinations of the current prevalence of alleles known to be associated with chloroquine, SP and other antifolates and mefloquine are in progress in many regions. For these drugs, we need to establish a common

Corresponding author: Sibley, C.H. (sibley@u.washington.edu).

Available online 24 April 2006

1471-4922/© 2006 Elsevier Ltd. All rights reserved.