# The promise of *T. cruzi* genomics

The publishing of the first *Trypanosoma cruzi* genome sequence was hailed as "a huge intellectual triumph", but what has it delivered?

It was 7 years in the making. Björn Andersson at the Karolinska Institute in Stockholm, Sweden, helped pioneer the *Trypanosoma cruzi* genome project from the first pilot sequencing project in 1998 to producing the first complete genome sequence in 2005 in collaboration with Najib El-Sayed at The Institute for Genomic Research (TIGR; now part of the J. Craig Venter Institute), and Ken Stuart and Peter Myler at the Seattle Biochemical Research Institute in the United States[1]. It was published alongside the sequences of two related species of parasite, which cause visceral leishmaniasis (*Leishmania major*) and sleeping sickness (*Trypanosoma brucei*). Together these three parasites are known as the TriTryps.

Of the known *T. cruzi* strains, CL–Brener was selected for sequencing as it was already well-studied experimentally. The sequencing effort revealed the assembled diploid genome to be 60.4 Mb in size, containing >22,000 protein-encoding genes, including a novel superfamily that codes for surface mucins, which might be crucial in helping the parasite evade the immune system. The project also revealed 6,200 'core proteins' present in all the TriTryps, which represent potential targets for broad-spectrum drugs. However, as many as 50% of the genes code for as-yet unknown proteins, hinting at entirely new areas for research[1].

## Taking the strain

*T. cruzi* strains are broadly classified into six lineages or discrete typing units (DTUs). Until recently these were divided into TcI and TcIIa–e, but they have now been re-designated as six separate groups TcI to TcVI, with different but not entirely exclusive geographical, ecological, transmission cycle and disease associations. The first genome sequence strain CL–Brener is a natural hybrid of TcII and TcIII. For comparison, Bjorn Andersson and his team have sequenced strain SylvioX10, representing the distinct TcI lineage, and compared it to CL–Brener, revealing that the two strains are more closely related than was previously thought. A full description of the SylvioX10-1 genome sequence will be submitted for publication shortly, to be followed by expression studies and protein analysis. Andersson's team and other groups also plan to sequence new *T. cruzi* genomes for 2011. Such comparative sequencing will contribute to epidemiological comparisons and the development of new lineage-specific diagnostic tests, and give further insight into the genetics of *T. cruzi*.

The sequence data also allowed researchers to probe the evolutionary origins of the trypanosomes, in particular the hypothesis that they descended from an ancestor that carried a photosynthetic intracellular symbiont — a plant-like parasite related to green algae. The sequences contained no plant-specific domains, thus the hypothesis was rejected[2]. There was a surprise in the RNA too: only *Trypanosoma brucei* appears to contain the machinery needed for RNA interference — a mechanism found in most eukaryotes. *T. cruzi* and most *Leishmania* appear to lack the enzymes required, suggesting that they might have other novel mechanisms for RNA regulation[2].

Besides providing new insights into the basic biology and evolution of *T. cruzi*, the genome sequence has allowed researchers to develop and refine a set of new tools with which to investigate parasite ecology, epidemiology and disease pathogenesis, as well as to enhance research and development of new drugs and diagnostics.

## From genes to drugs

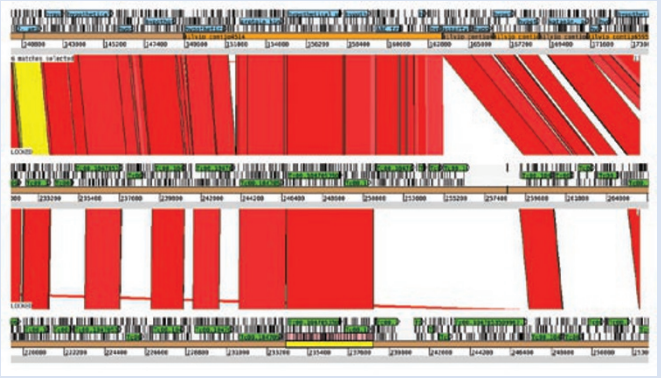
Sharing genomic information should speed up the drug-discovery efforts, which have lagged in recent decades (see Chagas disease: pushing through the pipeline on page S12). In particular, the World Health Organization (WHO)–Special Programme for Research and Training in Tropical Diseases (TDR) Targets database (Box 1) is an easy-to-use tool that allows researchers to select putative target proteins, ranked in terms of use as diagnostic markers

## Box 1 | Making sense of the data

New databases have been established to help researchers best use the burgeoning *Trypanosoma cruzi* genomic data. One is the collaborative *Trypanosoma brucei*, *T. cruzi* and *Leishmania* database (TriTrypDB; tritrypdb.org), which was set up by David Roos at the University of Pennsylvania in the United States with funding from the Bill & Melinda Gates Foundation. It integrates genome-scale datasets with information from functional genomics, to allow comparative analysis of the parasites.

Genomic datasets for a larger range of eukaryotic pathogens, including the TriTryps, and the genera *Plasmodium* and *Toxoplasma*, are combined at the Eukaryotic Pathogen Database (EuPathDB; eupathdb.org) created by the Eukaryotic Pathogen Bioinformatics Resource Center to allow more complex filtering of genetic and proteomic information.

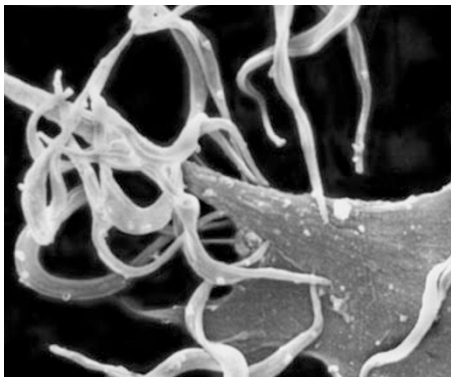
*T. cruzi* genome information is also held at the World Health Organization (WHO)–Special Programme for Research and Training in Tropical Diseases (TDR) Targets database (tdrtargets.org). This open-access database currently houses 11 genome sequences belonging to neglected tropical disease parasites, including helminths and those causing schistosomiasis, filariasis, malaria, tuberculosis and leprosy, together with related information on protein structure and function, and biochemistry and pharmacology, and is specifically designed for drug development.



OSCAR FRANZÉN

**Screen shots from a browser comparing parasite genomes.**

HELENE SANTOS BARBOSA



EYE OF SCIENCE/SCIENCE PHOTO LIBRARY

**The TriTryps. *Trypanosoma cruzi* trypomastigotes invade a cardiac myocyte (left), *Trypanosoma brucei* protozoa among red blood cells (middle) and *Leishmania major* parasitic protozoa (right).**

or druggable targets. Previously, target identification was a long and laborious process. "The goal of this project is to lower the energy barrier for doing this — it could take 1 or 2 years for a student to [identify drug targets] manually. The database allows researchers to rank and prioritize genes based on fundamental genomics or annotation or other sources of information", says the TDR Targets database leader Fernan Aguero who is at the Instituto de Investigaciones Biotecnologicas, Universidad Nacional de General San Martin, Buenos Aires, Argentina.

The database also includes a map of all sequence variation (single-nucleotide polymorphisms or SNPs) in *T. cruzi*, which could help identify different strains and, for drug development, indicate the risk of developing resistance. The next addition will be potential inhibitors, based on the European Bioinformatics Institute (EBI)'s ChEMBL drug-discovery database known as StARlite, to improve the drug-development process.

This kind of progress is satisfying for Andersson, whose initial motivation for producing a whole-genome sequence for *T. cruzi* was to help speed up target discovery. "The efforts underway to develop drug targets were very slow. For every project you had to identify a gene, clone it, characterize it, and do all the basic molecular biology work for any potential drug target. This would take years. With the genome you can look it up on the database and it will take you just 5 minutes. It has accelerated the entire field enormously."

## Investigating *T. cruzi* in the wild

Outside of drug development, the *T. cruzi* genome sequence is a major boost for epidemiologists wishing to track the geographical distribution of strains across South America, and to understand better their transmission, role in disease pathogenesis and susceptibility or resistance to drugs.

Such is the work undertaken by the European Union-funded network, ChagasEpiNet, which is coordinated by Michael Miles and Martin Llewellyn at the London School of Hygiene and Tropical Medicine in the United Kingdom. ChagasEpiNet links 15 research teams across South America and Europe, and uses genomic information not only to refine the mapping of *T. cruzi* strains but also to untangle some intriguing observations concerning the symptoms of Chagas disease, which appear to vary across South America. Geographically, TcI predominates north of the Amazon, whereas strains TcII, V and VI tend to occur more often south of the Amazon in Central and Eastern Brazil and the Southern Cone region (see Who, how, what and where? on page S8). This distribution seems to coincide with different chronic disease symptoms: all strains cause heart disease, but only the TCII, V and VI strains appear to be associated with chronic syndromes of the colon and oesophagus. Equally intriguing are reports that TcV might be easily transmitted congenitally — from mother to child — and that Chagas disease is easier to treat in some places than in others.

Miles suspects that this variation at least partially represents differences in the properties of the *T. cruzi* strains. "Different genotypes of *T. cruzi* are more distinct than named species of *Leishmania* [...] So it's inconceivable that they should all behave precisely the same. There is a lot of [experimental] evidence of different virulence and pathogenesis but it's not tied yet to any special molecular markers; there are no genetic determinants that mean we can say a particular lineage is more pathogenic, and no individual genes have been linked with severity of symptoms — yet — but that should come out of comparative genomics."

## Good-looking future

Now the first genome is complete, follow-on research projects are well underway: for example, proteomic studies are attempting to identify those proteins coded for by the mysterious 50% of the genome. Rick Tarleton at the University of Georgia, Athens, in the United States, is hoping this will reveal new protein targets, which in turn could serve as biomarkers for indicating disease course, and form the basis of new diagnostic tests. "We're cloning recombinant proteins of *T. cruzi* based on the genome sequence — avoiding any genes predominantly expressed in the insect vector and selecting those expressed in the human host. Using these we can look for changes in the T-cell responses over time upon drug treatment", says Tarleton

Andersson also has a new collaboration with Karsten Daub at the RIKEN Institute in Yoshihida, Japan, to isolate and sequence short RNAs of *T. cruzi* in the hope of identifying the novel RNA-regulatory mechanisms. They have so far found that particular groups of short RNAs cluster to certain gene families, and could therefore be involved in their regulation. "If we do find one of these RNAs and can show that it's regulatory, the mechanism is most likely to be a new one that nobody has seen before", he predicts.

These efforts — and research into *T. cruzi* epidemiology and pathogenesis — will be boosted by the sequencing of more strains, which is becoming increasingly cheaper and quicker to do. Sanger sequencing was used for the CL–Brener strain, which achieved 100–400 sequences (a sequence being 600–700 bases) per run. With the newer, faster techniques, it is possible to perform hundreds of thousands of sequences per run. "It's a totally different world now. We can sequence a trypanosome genome in a few days compared to a few years", says Andersson. This is all thanks to technological advances driven by the Human Genome Project. "The driver now is the challenge of the US$1,000 genome. It costs US$20,000 now, but I'm optimistic that sequencing is getting cheaper and soon it will be possible to fund sequencing through regular grants. The future looks good". ■

**Julie Clayton**

1. El-Sayed, N. M. *Science* **309**, 409–415.
2. El-Sayed, N. M. *Science* **309**, 404–409.