INSTITUT PASTEUR

Research in Microbiology 162 (2011) 619–625

www.elsevier.com/locate/resmic

The *Trypanosoma cruzi* genome; conserved core genes and extremely variable surface molecule families

Björn Andersson*

Department of Cell and Molecular Biology, Science for Life Laboratory, Karolinska Institutet, Berzelius väg 35, 171 77 Stockholm, Sweden

Received 27 February 2011; accepted 14 April 2011
Available online 18 May 2011

Abstract

The protozoan parasite *Trypanosoma cruzi* is an important but neglected pathogen that causes chagas disease, which affects millions of people, mainly in latin America. The population structure and epidemiology of the parasite are complex, with much variability among strains. The genome sequence of a reference strain, CL Brener, was published in 2005, and the availability of this sequence has both revealed the complexity of the parasite genome and greatly facilitated research into parasite biology and pathogenesis, by making the sequences of more than 8000 core genes available. The *T. cruzi* genome is highly repetitive, which has resulted in inaccuracies in the genome sequence, and attempts have been made to provide a deeper analysis of repeated genes as a complement to the genome sequence. The genome was found to be organized in stable core regions containing housekeeping and other genes, surrounded by highly repetitive, often sub-telomeric highly variable regions containing multiple members of large families of surface molecule genes. Comparative sequencing of *T. cruzi* strains has been initiated and the results show that the core gene content of the parasite is highly conserved, but that much sequence variability, 3–4% difference at the DNA level on average between strains in coding regions, is present. The additional genomes will improve the understanding of parasite biology and epidemiology.

Keywords: *Trypanosoma cruzi*; Genome; Repeated genes; Core gene set; Surface molecules

1. Introduction

The protozoan parasite *Trypanosoma cruzi* is the causative agent of chagas disease, which is a debilitating disease that infects at least 7.7 million people in latin America and causes 12,500 deaths annually (Rassi and Marin-Neto, 2010). Parasites are mostly transmitted when infected feces of the hematophagous triatomine insect vector make contact with mucosae or abraded skin after a blood meal. The parasites infect many different mammals and there are several different vector species, some of which are often found in houses and huts where people live. Other routes of transmission include blood transfusions and transmissions from mother to child during childbirth. The disease goes through an acute stage that is fatal in many cases. However, most morbidity is associated with the chronic stage of the disease, which can take several years to develop. The latter phase often includes heart symptoms, but also megasyndromes involving the gastrointestinal tract. There is no vaccine against *T. cruzi* infections and drug treatment is restricted to a small number of drugs with insufficient efficacy and potentially harmful side effects.

2. Population structure

T. cruzi is a diverse species. Multiple diverse strains with a high degree of variation in both karyotype, genome size and DNA sequence, as indicated by earlier comparative studies (Ruiz et al., 1998; Ramirez et al., 2010; Miles et al., 1981; Lewis et al., 2009), distributed over a large geographic area, as well as multiple insect and mammal hosts, have been described. Initial multilocus enzyme electrophoresis (MLEE), random amplification of polymorphic DNA (RAPD) and

* Tel.: +46 8 5248398; fax: +46 8 323950.
E-mail address: bjorn.andersson@ki.se.

0923-2508/$ - see front matter © 2011 Institut Pasteur. Published by Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.resmic.2011.05.003

microsatellite analyses suggested that the population structure of *T. cruzi* is predominantly clonal (Revollo et al., 1998, de Souza et al., 2003). However, there were also indications that genetic exchange contributed to the generation of strain diversity (Bogliolo et al., 1996). In a study, in which different regions of the genome in various isolates were compared, the existence of several hybrid lines was revealed (Augusto-Pinto et al., 2003). The mechanism of hybrid formation is not clearly understood. The current hypothesis is that cell fusion occurs at the intracellular stage, followed by a reduction in chromosome number accompanied by extensive recombination (Telleria et al., 2010).

The latest genotyping strategies have resulted in subdivision of *T. cruzi* into six phylogenetic groups, named discrete typing units (DTUs) I–VI (Zingales et al., 2009), where TcV and VI appear to be recent hybrids. Different DTUs are more prevalent in certain geographic locations and differences involving vectors, other hosts and environments, have also been suggested (Miles et al., 2009). For example, TcI is most common in chagas patients north of the amazon, e.g. (Ramirez et al., 2010; Miles et al., 1981; Sanchez-Guillem Ndel et al., 2006), but it also occurs in other regions (e.g. 14). TcII, TcV and TcVI are more common in the Southern Cone countries (Miles et al., 1981, 1984; Sanchez-Guillem Ndel et al., 2006; Burgos et al., 2010; Chapman et al., 1984; Barnabe et al., 2000, 2001; Bosseno et al., 2002; Breniere et al., 2002; Virreira et al., 2006; Cardinal et al., 2008). In addition, TcV and VI appear to be more closely associated with domestic transmission, i.e. with insect vectors that colonize human dwellings and transmit parasites to domestic animals and humans, compared to other strains. These conclusions are preliminary, as sampling is still limited and much work remains. The strains are not only genetically different. They also appear to vary in terms of physiology, biochemistry and infectivity (Ruiz et al., 1998; Revollo et al., 1998; de Souza et al., 2003; Augusto-Pinto et al., 2003; Zingales et al., 1999; Machado et al., 2006; Engel et al., 1990; Dvorak et al., 1982; Telleria et al., 2004) as indicated by studies of parasites in the laboratory. As the characterization of the genomic diversity of *T. cruzi* is still limited, it remains to be seen how all these differences affect pathogenicity and the highly variable clinical presentation of chagas disease.

3. Karyotype

The *T. cruzi* karyotype is complicated, with a relatively large number of chromosomes and tremendous size differences between homologous chromosomes both within and between strains. *T. cruzi* chromosomes do not condense during mitosis and it was therefore necessary to use alternative methods, mainly Southern blots of pulsed-field gel electrophoresis (PFGE) separated chromosomes followed by marker hybridization. Attempts to determine the sizes and numbers of chromosome in this way were carried out years before genome sequencing was initiated (Gibson and Miles, 1986; Engman et al., 1987; Krieger et al., 1990; Dietrich et al., 1990). The total DNA content was estimated to be between 90 and 150 Mb, for the mainly diploid genome, with estimates of the chromosome number varying from 64 to 80 chromosomes (Cano et al., 1995; Henriksson et al., 1996; Hanke et al., 1996). Several mapping studies were performed using small numbers of chromosome markers (Cano et al., 1995; Henriksson et al., 1996, 1993; Hanke et al., 1996), but the large number of co-migrating chromosomes, and the fact that only a small number of genetic markers were available made it impossible to determine an accurate karyotype. However, a partial *T. cruzi* karyotype began to emerge. A later larger study of the CL Brener strain (Porcile et al., 2003) used more chromosome-specific markers and several repetitive elements, but did not reveal the number and size of all chromosomes. As part of the *T. cruzi* genome project, we made an attempt to produce a complete karyotype of three strains of *T. cruzi,* including the genome reference strain CL Brener Sylvio X10/7 and CAI/72 (Branche et al., 2006). A set of 239 cDNA markers were hybridized to Southern blots of chromosomes separated using three different PFGE conditions. The larger number of markers compared to previous studies made it possible to identify fifteen linkage groups and the number of chromosomes was estimated at 55 to 57 in total. Data confirmed the enormous variation in chromosome size between strains and between homologous chromosomes within a strain. Several differences larger than a megabase were found. While these and results from other chromosomes improved greatly on the available karyotype, chromosome numbers were still not certain and marker data were not used extensively for finishing of the *T. cruzi* CL Brener genome. The best estimate to date of the karyotype was produced using sequence and scaffolding data from the *T. cruzi* genome project together with markers to identify longer scaffolds (Weatherly et al., 2009). This study estimated that the *T. cruzi* genome is organized into 41 pairs of chromosomes, 82 in total, a larger number than previously thought. However, the exact number is still uncertain and may change.

4. The reference genome – CL Brener

The *T. cruzi* genome project began in 1990 as a WHO/TDR-funded network of laboratories that began to accumulate resources, such as clone libraries and EST sequences and a few initial shorter genome sequences (Andersson et al., 1998), for the mapping and sequencing of the selected reference strain CL Brener, from the TcVI Distinct Typing Unit (DTU). This strain was selected based on the clinical picture and laboratory characteristics, and not much was known regarding the characteristics of this particular genome at the time. In 2000, the laboratories of Najib El-Sayed (TIGR), Ken Stuart and Peter Myler (SBRI) and my laboratory initiated an NIH-funded project to completely sequence the CL Brener genome. The project began using a clone-by-clone strategy, but due to uneven BAC-clone coverage, it was decided to switch to a whole genome shotgun strategy. In this way, more than 1.2 million Sanger sequence reads were generated from several libraries with different insert sizes. The resulting whole genome shotgun assembly was relatively fragmented because