Parasite Genome Projects and the *Trypanosoma cruzi* Genome Initiative

Wim Degrave/+, Mariano J Levin*, José Franco da Silveira, Carlos M Morel**

Departamento de Bioquímica e Biologia Molecular, Instituto Oswaldo Cruz, Av. Brasil 4365, 21045-900 Rio de Janeiro, RJ, Brasil *Instituto de Investigaciones en Ingenieria Genetica y Biologia Molecular, INGEBI-CONICET, FCEyN-UBA, Obligado 2490, 1428 Buenos Aires, Argentina **Universidade Federal de São Paulo (UNIFESP-EPM), Rua Botucatu 862, 04023-062 São Paulo, SP, Brasil

Since the start of the human genome project, a great number of genome projects on other "model" organism have been initiated, some of them already completed. Several initiatives have also been started on parasite genomes, mainly through support from WHO/TDR, involving North-South and South-South collaborations, and great hopes are vested in that these initiatives will lead to new tools for disease control and prevention, as well as to the establishment of genomic research technology in developing countries. The Trypanosoma cruzi *genome project, using the clone CL-Brener as starting point, has made considerable progress through the concerted action of more than 20 laboratories, most of them in the South. A brief overview of the current state of the project is given.*

Key words: *Trypanosoma cruzi* - clone CL Brener - Parasite Genome Projects

The determination of the complete nucleotide sequence of an organism has attracted researchers ever since there were technical means to do it. The first virus to be sequenced, using enzymatic RNA sequencing, was MS2 (Fiers et al. 1976), followed by bacteriophage ΦX (Sanger et al. 1977a) using his enzymatic method (Sanger et al. 1977b) and SV40 (Fiers et al. 1978), using Maxam and Gilbert's (1977) chemical degradation method. The human mitochondrial genome (Anderson et al. 1981), the lambda genome (Sanger et al. 1982) and bacteriophage T7 (Dunn & Studier 1983) were other important marks, and much was learned about the lifecycle of viruses and their interaction with the infected host cell. Indeed, a considerable number of viral genomes have been determined since then, both from plant, animal and human infections. However, sequencing was quite tedious, and only in the late eighties, daring plans were made to launch a massive attack on the human genome, with the advent of automated sequencing. Much has happened since its official start in 1990, and today, hardly any scientist will ignore, let alone deny the usefulness and technical feasibility of the human genome project. Considerable technical progress has been made in large scale mapping and sequencing, and the cost and speed of the project is improving faster than anticipated. Quantum leap improvements in sequencing are being projected for the coming years (ex. Burke et al. 1997), which will transform molecular biology into a discipline of bioinformatics, study of structure-function (Bassett et al. 1996, Oliver 1996), and manipulation of living organisms, rather than the analytical science it has been over the last 15 years.

The start of a great number of genome projects on other organisms than human has been a remarkable, and at first unexpected phenomenon. On the one hand they figured as an "exercise" on smaller genomes, but undoubtedly they also came about due to the enthusiasm of molecular biologists, once it was realized that (the human) genome projects were not a crazy undertaking (Tilghman 1996). Thus, the complete genomes of *Haemophilus influenzae* Rd, *Mycoplasma genitalium*, *Methanococcus jannaschii*, *Synechocystis* sp. (PCC 6803*)*, *Mycoplasma pneumoniae*, *Saccharomyces cerevisiae* and *Escherichia coli* K-12 are now known, and a great number of other genome sequences are being determined, mainly from bacteria, but also from many higher eukaryotes, including plant and animal "model organisms". We can thus expect that our understanding of the complete gene set of many different life forms will increase dramatically over the coming years, and that regulation of gene expression, short and long range genomic DNA interactions, interaction of organisms with their environment, genome evolution,

Financial support: UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases; CYTED- Subprogram of Biotechnology (Spain), CNPq, Centro Brasileiro Argentino de Biotecnologia (CBAB/CABBIO).

+Corresponding author: Fax: +55-21-590.3495. E-mail: wdegrave@gene.dbbm.fiocruz.br

Received 20 August 1997
Accepted 10 September 1997

embryogenesis, interactions between cells in a multi-cellular organism and mechanisms of molecular evolution, amongst other aspects of cellular and molecular biology, will come under a completely new light.

As was pointed out by Pena (1996), countries in the developing world have not been able to contribute significantly to the aforementioned genome projects, although their population is in great need for the benefits from the expected outcome of such projects: they host more than 70% of the world's population, struggle with major public health problems caused by infectious and parasitic diseases, suffer from food shortage, and face a further increasing technology gap.

Partly in response to the above picture, scientists from developing and developed countries planned and initiated a number of parasite genome projects and several "consortiums" for the mapping and sequencing of these medium-sized genomes were established, often based on already ongoing scientific collaborations. Thus, the genomes of *Plasmodium falciparum*, *Schistosoma mansoni*, *Trypanosoma cruzi*, *Leishmania major*, *Trypanosoma brucei*, *Brugia malayi* are now under study (Zingales et al. 1997a). The main objectives for these projects are the following: (1) increase drastically the knowledge on the (molecular) biology of these parasites, which are, from an evolutionary standpoint, very ancient and present a number of quite unique features and molecular processes; (2) identify, as fast as possible, new genes with key cellular functions, which could be eligible as target for new drugs; (3) identify new antigens which could be useful in diagnostics or for vaccine development; (4) build expertise and collaborative North-South and South-South networks for genome research in the fields of mapping, large scale sequencing, bioinformatics, research on protein structure-function relationships etc.; (5) contribute to the overall knowledge on genome structure and comparative biology/evolution.

In this volume, several papers were brought together, as a result of the International Training Course on Parasite Genome Projects: Strategies and Methods, and the Symposium on Genome Projects, held at INGEBI, Buenos Aires, Argentina, Nov. 13-24, 1995, dealing mainly with aspects of the *T. cruzi* genome project, amongst contributions from other parasite genome projects (Musto et al. 1997, Requena et al. 1997, Tanaka et al. 1997).

THE *TRYPANOSOMA CRUZI* GENOME PROJECT

The idea of a *Trypanosoma cruzi* Genome Project began to take form during several independent meetings in 1993 and 1994, especially the Annual Meeting for Basic Research in Chagas Disease (Caxambu, Brazil, 5-10 November 1993), the French-Latin American Meeting at INGEBI (Buenos Aires, Argentina, 24-26 November 1993) and the Annual Meeting of the CYTED-D Biotechnology Program, at the University of Chile (Santiago, Chile, March 1994) and was finally crystalized at the WHO/TDR-Fiocruz co-sponsored Parasite Genome Network Planning Meeting at Fiocruz (Rio de Janeiro, Brazil, 14-15 April 1994), where a *T. cruzi* workgroup of about 15 researchers from different countries in the Americas decided to go ahead with the project in a concerted effort, choosing as starting point a *T. cruzi* clone from CL strain, now called CL-Brener.

The first meeting of the *T. cruzi* Genome Initiative, held in Teresópolis (Brazil, April 10-12, 1996), sponsored by WHO-TDR, and with cosponsoring by Fiocruz, CYTED-D, the French-Latin American Cooperation and the University of Uppsala, Sweden, brought together 20 researchers and coordinators from 18 different laboratories in 9 countries, to discuss the first progress report and future planning of the project. The second meeting was held in Rio de Janeiro (Brazil, June 9-11 1997) under the same framework.

The WHO-TDR network, with very limited funding considered as "seed money", was started in 1994, and the parasite genome initiatives are now very well developed in de participating developing countries, with the expectation that genome projects will provide new answers and tools for disease control, as the current tools are not enough. Although Chagas disease transmission is now under control in some countries in Latin-America, the *T. cruzi* genome initiative is still very much needed in order to understand more about Chagas disease and the parasite-host interactions, to develop new drugs for efficient treatment, and to allow for comparative biology studies with *Leishmania* and *T. brucei*. The current state of the project can be summarized as follows.

Characterization of T. cruzi CL-Brener (Zingales et al. 1997b) - The biological and parasitological characteristics of the clone CL-Brener have been carefully studied, as well as molecular typing of CL-Brener, using zymodeme, schizodeme, RAPD and molecular markers, including the use of such markers for the testing of the genetic stability of CL-Brener. No variation was seen after 100 generations and no significant differences with the parental CL strain have been reported up to now. The Reference Laboratory (USP, São Paulo), charged with the distribution of CL-Brener, has sent the clone to many laboratories in 11 countries. The recent discovery of at least two main lineages amongst *T. cruzi* strains (Souto et al. 1996) was discussed, however it was not felt that this

would affect the genome project. A representative of "group 2", clone Dm28c, was chosen for analysis and comparison with CL-Brener, which belongs to "group 1".

The molecular karyotype of CL-Brener (Santos et al. 1997) - Karyotype analysis under different resolutions, chromosomal marker production and linkage analyses have been carried out by different groups and, depending on pulse field gel electrophoresis conditions (Hernandez-Rivas et al. 1997), about 20 to 42 chromosomal bands, ranging from 0.45 Mb to 3.5 Mb, can be distinguished, numbered as I to XX (low resolution) or 1 to 42 (medium resolution) (Henriksson et al. 1995). However, the real number of chromosomal bands is undoubtedly higher, and a total genome content of about 87 Mb can be expected. Several linkage groups have been identified and *T. cruzi* is believed to be diploid, while several apparently homologous chromosomes have different sizes with sometimes large differences (ex. 580 kb versus 800 kb). Only few (50) gene probes had been assigned thus far to chromosomal bands. Karyotype pattern comparisons between different *T. cruzi* strains/clones showed substantially different patterns in each case. It was concluded that there is thus far no reason to believe that CL-Brener would have a substantially larger genome size than other *T. cruzi* strains, nor that its karyotype pattern would be more difficult to analyze.

Genomic and cDNA library construction of CL-Brener - CL-Brener YAC, BAC (Ferrari et al. 1997) and cosmid libraries (Hanke et al. 1996), are available from the respective laboratories. However, a major effort to grid the YAC and BAC libraries on (high density) filters is called for. Moreover, the construction of a new, high quality BAC library is highly desirable and will proceed with the support of CEPH (Paris, France). The start of contig construction, using a gridded cosmid library, was reported for chromosomes 1, 3 and 4.

An epimastigote cDNA library was constructed by directional cloning, as well as a normalized cDNA library.

EST and STS sequencing efforts - In total, 265 EST sequences were obtained from the non-normalized library, and up to now, 600 from the normalized library at Fiocruz (Brandão et al. 1997), 1300 in Uppsala and 90 at IPB, Grananda, Spain, totaling around 2300 sequences. Also, new repetitive sequences, such as SIRE, Viper etc., were characterized, as well as minisatellite sequences and STS, mainly obtained by inter-SIRE PCR. An effort is now being made to map these sequences onto the chromosomal map. A standardized nomenclature for EST and STS clones was accepted.

Genomic sequencing - The Swedish group (Uppsala, Sweden) reported the sequence of one complete cosmid, 5 near to complete cosmids and 3 in the final phase, totaling about 300 kbp from chromosome 3. Thirty six genes were identified. Contig mapping of chromosome 1-4 is under way, preceding the sequencing phase.

Data analysis, database construction and distribution - Although sequence deposit during this first sequencing phase has not been immediate, it was agreed that all sequences will be displayed on the WWW, and that data deposit to Genbank/EMBL should be immediate. Incorporation into TcruziDB, under construction at Fiocruz, Brazil and available via ftp (Degrave et al. 1997), will be done after deposit and data analysis. With the availability of "winace", the win95/NT version of ACeDB, access to TcruziDB will be greatly improved. Furthermore, two discussion lists are available for communication and coordination: tcruzi-l@iris.dbbm.fiocruz.br and tcgenics@iris.dbbm.fiocruz.br (Degrave et al. 1997).

THE POST-GENOME INITIATIVE

Under the auspices of WHO/TDR, a post-genome initiative is now under discussion. Post-genome efforts are in the first place concerned with applications of the genome projects for disease control and prevention, leading to new drugs, vaccines and other products. A close collaboration with industry seems preemptive. However, several difficulties can be foreseen, such as the still fragile structure of the parasite genome "consortiums" in this initial stage of the large scale sequencing phase, the current lack of substantial financial resources both for the genome projects in itself and for any new phase, the delicate interactions between basic research and industrial applications (research laboratories versus industrial partners), the need for sophisticated studies on protein structure function relationships etc. More than ever, North-South collaborations will play a key role in these discussions.

ACKNOWLEDGMENTS

To Drs J Swindle, B Dobrokhotov, B Andersson, J Kelly, U Pettersson, E Rondinelli, J Bua, A Ruiz, A Macedo, B Zingales, C Frasch, D le Paslier, J Ramirez, R Aldao, A Gonzalez, A De Miranda, A Brandão, O Fernandes, J Requena, and other collaborators in the *T. cruzi* genome initiative for their participation and contributions during the *T. cruzi* genome network planning meetings.

REFERENCES

Anderson S, Bankier AT, Barrel BG, de Brujin MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R,

Young IG 1981. Sequence and organization of the human mitochondrial genome. *Nature 290*: 457-465.

Basset Jr DE, Boguski MS, Heiter P 1996. Yeast genes and human disease. *Nature 379*: 589-590.

Brandão A, Urmenyi T, Rondinelli E, de Miranda AB, Gonzalez A, Degrave W 1997. Identification of transcribed sequences (ESTs) in the *Trypanosoma cruzi* genome project. *Mem Inst Oswaldo Cruz* (this volume).

Burke DT, Burns MA, Mastrangelo C 1997. Microfabrication technologies for integrated nucleic acid analysis. *Genome Res 7*: 189-197.

Degrave W, de Miranda AB, Amorim A, Brandão A, Aslett M, Vandeyar M 1997. TcruziDB, an integrated database, and the WWW information server for the *Trypanosoma cruzi* genome project. *Mem Inst Oswaldo Cruz* (this volume).

Dunn JJ, Studier FW 1983. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J Mol Biol 166*: 477-535.

Ferrari I, Lorenzi H, Santos MR, Brandariz S, Requena JM, Schijman A, Vázquez M, da Silveira JF, Ben-Dov C, Medrano C, Ghío S, Bergami PL, Cano I, Zingales B, Urmenyi TP, Rondinelli E, González A, Cortes A, Lopes MC, Thomas A, Alonso C, Ramírez JL, Chiurrillo MA, Aldao RR, Brandão A, Degrave W, Perrot V, Saumier M, Billaut A, Cohen D, Le Paslier D, Levin MJ 1997. Towards the physical map of the *Trypanosoma cruzi* nuclear genome: construction of YAC and BAC libraries of the reference clone *T. cruzi* CL Brener. *Mem Inst Oswaldo Cruz* (this volume).

Fiers W, Contreras R, Duerinck F, Haegemann G, Iserentant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers A, Van den Berghe A, Volckaert G, Ysebaert M 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature 260*: 500-507.

Fiers W, Contreras R, Haegemann G, Rogiers R, Van de Voorde A, Van Heuversweyn H, Van Herreweghe J, Volckaert G, Ysebaert M 1978. Complete nucleotide sequence of SV40 DNA. *Nature 273*:113-120.

Hanke J, Sánchez DO, Henriksson J, Aslund L, Pettersson U, Frasch ACC, Hoheisel JD 1996. Mapping the *Trypanosoma cruzi* genome: analyses of representative cosmid libraries. *Biotechniques 21*: 686-693.

Henriksson J, Porcel B, Rydaker M, Ruiz A, Sabaj V, Galanti N, Cazzulo JJ, Frasch AC, Pettersson U 1995. Chromosome specific markers reveal conserved linkage groups in spite of extensive chromosomal size variation in *Trypanosoma cruzi*. *Mol Biochem Parasitol 73*: 63-74.

Hernandez-Rivas R, Scherf A 1997. Separation and mapping of chromosomes of parasitic protozoa. *Mem Inst Oswaldo Cruz* (this volume).

Maxam A, Gilbert W 1977. A new method for sequencing DNA. *Proc Natl Acad Sci USA 74*: 560-564.

Musto H, Cacciò S, Rodríguez-Maseda H, Bernardi G 1997. Compositional constraints in the extremely GC-poor genome of *Plasmodium falciparum*. *Mem Inst Oswaldo Cruz* (this volume).

Oliver SG 1996. From DNA sequence to biological function. *Nature 379*: 597-600.

Pena SDJ 1996. Third world participation in genome projects. *Tibtech 14*:74-77.

Requena JM, Soto M, Quijada L, Alonso C 1997. Genes and chromosomes of *Leishmania infantum*. *Mem Inst Oswaldo Cruz* (this volume).

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB 1982. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol 162*: 729-773.

Sanger F, Air GM, Barrel BG, Brown NL, Coulson AR, Fiddes JC, Hutchinson CA, Solcombe PM, Smith M 1977a. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature 265*: 687-695.

Sanger F, Nicklen S, Coulson AR 1977b. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA 74*: 5463-5467.

Santos MRM, Cano MI, Schijman A, Lorenzi H, Vázquez M, Levin MJ, Ramirez JL, Brandão A, Degrave WM, da Silveira JF 1997. The *Trypanosoma cruzi* genome project: nuclear karyotipe and gene mapping of clone CL Brener. *Mem Inst Oswaldo Cruz* (this volume).

Souto RP, Fernandes O, Macedo AM, Campbell DA, Zingales B 1996. DNA markers define two major phylogenetic lineages of *Trypanosoma cruzi*. *Mol Biochem Parasitol 83*: 141-152.

Tanaka M, Tanaka T, Inazawa J, Nagafuchi S, Kitamura L, Mutsui Y, Kaukas A, Johnston DA, Rollinson D 1997. Proceedings of the Schistosome genome project. *Mem Inst Oswaldo Cruz* (this volume).

Tilghman SM 1996. Lessons learned, promises kept: a biologist's eye view of the genome project. *Genome Res 6*: 773-780.

Zingales B, Pereira MES, Almeida KA, Umezawa ES, Nehme NS, Oliveira RP, Macedo A, Souto RP 1997a. Biological parameters and molecular markers of clone CL Brener - the reference organism of the *Trypanosoma cruzi* genome project. *Mem Inst Oswaldo Cruz* (this volume).

Zingales B, Rondinelli E, Degrave W, da Silveira JF, Levin M, Le Paslier D, Modabber F, Dobrokhotov B, Swindle J, Kelly JM, Aslund L, Hoheisel JD, Ruiz AM, Cazzulo JJ, Pettersson U, Frasch AC 1997b. The *Trypanosoma cruzi* genome initiative. *Parasitol Today 13*: 16-22.