Status of protozoan genome analysis: trypanosomatids

J. M. BLACKWELL[1,2] *and* S. E. MELVILLE[2]

[1] *Cambridge Institute of Medical Research, Wellcome Trust/MRC Building, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2XY*
[2] *Department of Pathology, Tennis Court Road, Cambridge CB2 1QP*

SUMMARY

The three trypanosomatid genome projects have employed common strategies which include: analysis of pulsed-field gel electrophoretic chromosomal karyotypes; physical mapping using big DNA (cosmid, pacmid P1, bacterial artificial chromosome, yeast artificial chromosome) libraries; partial cDNA sequence analysis to develop sets of expressed sequence tags (ESTs) for gene discovery and use as markers in physical mapping; genomic sequencing; dissemination of information through development of web-sites and ACeDB-based fully integrated databases; and establishment of functional genomics programmes to maximize useful application of genome data. Highlights of the projects to date have been the demonstration that, despite extensive chromosomal size polymorphisms for diploid homologues within Africa trypanosomes, *T. cruzi* or *Leishmania*, the physical linkage groups for markers on each chromosome are retained across all isolates/species studied within each group. For African trypanosomes, detailed analysis of chromosome 1 has demonstrated that repetitive sequences and the two retroposon-like elements RIME and INGI are localized to a defined region at one end of the chromosome, with the bulk of the central region of the chromosome containing genes coding for expressed proteins. Comparative mapping shows that, although subtelomeric changes account for a large proportion of the polymorphism in chromosome size in African trypanosomes, there are significant expansions and contractions in regions across the entire chromosome. The highlight of the genomic sequencing projects has been the demonstration of just 2 putative transcriptional units of chromosome 1 of *Leishmania major*, extending on opposite strands from a point in the central region of the chromosome. A similar observation made on 93·4 kb of contiguous sequence for *T. cruzi* chromosome 3 suggests the presence of promoter and regulatory elements at the junctions of large polycistronic transcriptional units. All data obtained from the genome projects are made available through the public domain, which has prompted changing philosophies in how we approach analysis of the biology of these organisms, and strategies that we can employ now in the search for new therapies and vaccines.

Key words: Trypanosomatids; genome analysis, karyotypes, physical mapping, sequencing, functional genomics.

INTRODUCTION

The 3 trypanosomatid genome projects have employed common strategies which include: analysis of pulsed-field gel electrophoretic (PFGE) chromosomal karyotypes; physical mapping using big DNA (cosmid, pacmid P1, bacterial artificial chromosome or BACs, yeast artificial chromosome or YAC) libraries; partial cDNA sequence analysis to develop sets of expressed sequence tags (ESTs) for gene discovery and use as markers in physical mapping; genomic sequencing; dissemination of information through development of web-sites and ACeDB-based fully integrated databases; and establishment of functional genomics programmes to maximize useful application of genome data (reviewed by Blackwell, 1997; de Grave *et al.* 1997; Melville, 1997, 1998; Melville *et al.* 1998*a*; Ivens & Blackwell, 1996, 1999). The ability to carry out experimental crosses has also allowed the African trypanosome genome project to work on the development of a genetic map for *Trypanosoma brucei*, which can be employed in positional cloning studies. This paper presents some of the highlights of the genome projects for the three different trypanosomatid parasites: *T. brucei*, *T. cruzi*, and *Leishmania major*.

THE AFRICAN TRYPANOSOME GENOME PROJECT

The TREU 927/4 strain of *T. brucei* was selected as the strain for genome analysis and sequencing (Melville, 1997, 1998). This was fortuitous since subsequent comparative PFGE karyotype analysis has demonstrated that this strain has the smallest genome size for the megabase chromosomes, 53·4 Mb compared to 67·3 Mb and 71·2 Mb for other well-characterized laboratory strains STIB 386 and STIB 247. The highlight of this karyotype analysis, which has involved establishing a nomenclature (Turner, Melville & Tait, 1997) and the mapping of $>$ 400 ESTs as markers onto the PFGE karyotype (Melville *et al.* 1998*b*), has been the demonstration that, despite the very dramatic chromosomal size polymorphisms for diploid homologues (up to 200% within and 400% between individual isolates), the physical linkage groups for markers on each chromosome are retained across all isolates studied and in the hybrid progeny of experimental crosses. This mapping study has drawn on probes from $>$ 3500 EST sequences now deposited in the public domain for *T. brucei*, the gene discovery element of which has already seeded many new research projects and provided a resource of clones accessed internationally.

Parasitology (1999), **118**, S11–S14. Printed in the United Kingdom
https://doi.org/10.1017/S0031182099004163 Published online by Cambridge University Press

Development of a detailed map across chromosome I (Melville, Gerrard & Blackwell, 1999), a megabase chromosome in strain TREU 927/4, has demonstrated that repetitive sequences and the 2 retroposon-like elements RIME and INGI are localized to a defined region at one end of the chromosome, and that the diploid pair of chromosome I in this strain has only one bloodstream form variant surface glycoprotein (VSG) expression site telomeric to the report sequences. A metacyclic VSG expression site occurs at the opposite telomere on 1 homologue, adjacent to isolated copies of RIME and INGI. The bulk of the central region of the chromosome, where genes coding for expressed proteins are located, contains no RIME or INGI sequences. Comparative mapping of this chromosome across different isolates has further demonstrated that, although subtelomeric changes account for a large proportion of the polymorphism in chromosome size, there are significant expansions and contractions in regions across the entire chromosome. Given that the physical linkage groups/gene order are maintained across the central region of the chromosome, it will be of some interest to determine the (?mobile) elements involved in expansion/contraction at different sites across the chromosome. This will be greatly facilitated by genomic sequence analysis ($\approx 5\times$ cover) of chromosome I from strain 927, currently in progress at the Sanger Centre (Hinxton, Cambridge UK).

At TIGR in the USA, a different strategy is being adopted for genomic sequence analysis, involving end-sequence analysis of approximately 20000 clones from a variety of libraries: cosmid, P1 and BAC. This will provide about 20 Mb of single pass, non-contiguous sequence, which will enhance gene discovery and provide markers for construction of a global physical map for strain TREU 927/4. This end-sequence analysis will be completed during 1999, allowing the *T. brucei* genome project to reassess strategies (e.g. whole genome shot-gun versus chromosome-by-chromosome approaches) to be used to obtain complete contiguous sequence coverage of the genome.

A first generation genetic map has been developed using 200 Amplified Fragment Length Polymorphic (AFLP) markers (Masiga, Turner and Tait, personal communication), demonstrating 11 linkage groups which match the 11 diploid pairs of megabase chromosomes identified by PFGE karyotype analysis (Melville *et al*. 1998*b*). Further development of the genetic map, including mapping of segregating phenotypes in the progeny of genetic crosses, together with the availability of the full genomic sequence, will make it possible to identify genes controlling important phenotypes (e.g. drug resistance, virulence determinants, etc.) by positional cloning.

Resources available from the African trypanosome genome project are detailed in Melville *et al*. (1998*a*). Details of the integrated database TrypDB are available at http://parsun1.path.cam.ac.uk.

THE *T. CRUZI* GENOME PROJECT

Clone CL Brener is the reference organism used in the *T. cruzi* genome project (Cano *et al*. 1995; Zingales *et al*. 1997). The estimated diploid genome size for this strain is ~ 87 Mb (Santos *et al*. 1997). Although chromosomal size polymorphism is not as dramatic as in African trypanosomes, it is sufficient to facilitate PFGE karyotype mapping by analysis of multi-strain blots (Henriksson *et al*. 1995, 1996). Chromosome-specific markers again reveal conserved linkage groups in spite of extensive chromosomal size variation in *T. cruzi*. Physical mapping projects are in progress (Ferrari *et al*. 1997; Frohme *et al*. 1998*a*, *b*), utilizing big DNA (cosmid, BAC and YAC) libraries made for the genome strain CL Brener. Through the efforts of the *T. cruzi* genome network (e.g. Verdun *et al*. 1998), >6500 EST sequences are now on the public domain, and genomic sequence analysis of the smallest (~ 670 kb) chromosome 3 of *T. cruzi* is almost complete. The first 93·4 kb of sequence contained 20–30 novel genes and several repeat elements, including a novel chromosome 3-specific 400-bp repeat sequence (Andersson *et al*. 1998). The intergenic sequences were found to be rich in di- and trinucleotide repeats of varying lengths and also contained several known *T. cruzi* repeat elements. An interesting feature of the sequence was that the genes appear to be organized in 2 long clusters containing multiple genes on the same strand. The 2 clusters are transcribed in opposite directions and they are separated by an ~ 20 kb GC-rich sequence containing 2 large repetitive elements, as well as a pseudogene for cruzipain and a gene for U2snRNA. The authors speculate that this strand switch region contains one or more regulatory and promoter regions. Sequencing a chromosome 4, and end-sequence analysis of the BAC library, is also in progress. Like the *T. brucei* project, this network will be reviewing the most appropriate strategy to pursue full genomic sequence analysis of *T. cruzi*.

Details of the integrated database, TcruziDB, and www server (http://www.dbbm.fiocruz.br) which support the *T. cruzi* genome project are provided in Degrave *et al*. (1997).

THE *LEISHMANIA* GENOME PROJECT

PFGE karyotype analysis, again relying on chromosomal size polymorphism across isolates and analysis of multi-species blots (Wincker *et al*. 1996), is complete for the genome strain *L. major* Friedlin (Bastien *et al*. 1998). The first generation cosmid physical map is also complete (Ivens *et al*. 1998), and

is being used to underpin a chromosome-by-chromosome approach to genomic sequence analysis. The full sequence for chromosomes 1 (250 kb) and 3 (350 kb) are complete, and significant progress has been made on chromosomes 2, 4, 5 and 6. Around 7·7 million dollars has been raised for genomic sequence analysis, which is being carried out in the USA (Seattle Biomedical Research Institute; NIH-funded), at the Sanger Centre (Hinxton, Cambridge, UK; Wellcome Trust funded), by a consortium of SMEs across the European Union (EU-funded), and in Brazil (Brazilian Government- and WHO-funded). Funding to hand will allow completion of 45 % of the genomic sequence during the year 2000. The nature of the funding, and the progress already made, means that the *Leishmania* genome network is almost committed to continuing the chromosome-by-chromosome sequencing strategy, using a combination of cosmid sequence analysis and whole chromosome shot-gun sequence analysis. The highlight of the genomic sequencing project has been the demonstration of just 2 putative transcriptional units on chromosome 1 extending on opposite strands from a point in the central region of the chromosome (Myler *et al*. 1999). This is reminiscent of the observation made with the *T. cruzi* contiguous sequence (Andersson *et al*. 1998), suggesting again the presence of promoter and regulatory elements at the junction of the two putative transcriptional units. Interestingly, chromosome 3 shows a similar organization, but with the 2 transcriptional units extending from the ends of the chromosome towards the middle (P. Myler, Seattle, USA, personal communication). One advantage of sequencing multiple trypanosomatids is that it will ultimately reveal the extent to which organization of genes is conserved across trypanosomatids. Some evidence for this is becoming available (Bringaud *et al*. 1998), and the identification of common polycistronic transcriptional units could have important implications for therapeutic targeting of related groups of genes.

Activities in the *Leishmania* genome network are now focusing on functional genomics. Programmes include: screening of microarrays for cDNA clones from the EST sequencing project, and ORFs from the genomic sequence, using mRNA from different developmental stages, strains of differing virulence, and different species of *Leishmania*; construction of a series of chromosomal deletions using chromosomal fragmentation techniques, with a view to generating a series of haploidized chromosomes to facilitate one-step knockout strategies; phenotypic and functional analysis of knockouts for the ORFs of chromosome 1 generated using a PCR-based strategy; systematic knockout of overlapping cosmids on chromosome 4 followed by phenotypic and functional analyses; and development of generic antisense strategies for analysis of gene function.

Details of resources available from, and the integrated database LeischDB for, the *Leshmania* genome project are at http://www.eb.ac.uk/parasites/leish.html.

CONCLUSIONS

Incredible progress has been made in analysis of trypanosomatid genomes, with the individual genome projects moving to genomic sequencing much faster than any of us had anticipated. This has led to changing philosophies in how we approach analysis of the biology of these organisms, and strategies that we can employ now in the search for new therapies and vaccines. In our laboratory we have been using a pooling strategy for DNA vaccine testing in mice to identify potential new vaccine candidates for trial in primates and dogs. The beauty of this approach is that we can test new molecules as vaccines without the need to make protein, or the need to know anything about the identity or function of the gene. This is fortuitous, since only about 30–50 % of the genes sequenced so far in the EST of genomic sequencing projects has identity to known proteins in the public databases. So far we have used cDNA clones from the EST project. However, the fact that leishmanial genes contain no introns, means that we can readily change our strategy to systematically test the PCR-amplified ORFs from each chromosome as the sequence data are released. This demonstrates one of the great things about genome projects – the public availability of the data – making it possible for scientists around the world to access and incorporate data on gene sequences and expression profiles into their own research programmes. The new millennium should see in a generation of scientists who no longer have to think about, or write grants to fund, the sequencing their gene(s) of interest. We can all focus our minds on new and imaginative ways to apply the genome data to the critical problems of disease control.

ACKNOWLEDGEMENT

The trypanosomatid genome projects have received core support from the UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (TDR).

REFERENCES

ANDERSSON, B., ASLUND, L., TAMMI, M., TRAN, A. N., HOHEISEL, J. D. & PETTERSSON, U. (1998). Complete sequence of a 93·4-kb contig from chromosome 3 of *Trypanosoma cruzi* containing a strand-switch region. *Genome Research* **8**, 809–816.

BASTEIN, P., BLAINEAU, C., BRITTO, C., DEDET, J.-P., DUBESSAY, P., PAGES, M., RAVEL, C., WINCKER, P., BLACKWELL, J. M., LEECH, V., LEVICK, M., NORRISH, A., IVENS, A., LEWIS, S., BAGHERZADEH, A., SMITH, D.,

MYLER, P., STUART, K., CRUZ, A., RUIZ, J. C., SCHNEIDER, H., SAMPAIO, I., ALMEDIA, R., PAPADOPOULOU, B., SHAPIRA, M., BELLI, S. & FASEL, N. (1998). The complete chromsomal organization of the reference strain of the *Leishmania* genome project, *L. major* 'Friedlin'. *Parasitology Today* **14**, 301–303.

BLACKWELL, J. M. (1997). Protozoan parasite genome analysis: progress in the African trypanosome and leishmanial genome networks. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **91**, 107–110.

BRINGAUD, F., VEDRENNE, C., CUVILLIER, A., PARZY, D., BALTZ, D., TETAUD, E., PAYS, E., VENEGAS, J., MERLIN, G. & BALTZ, T. (1998). Conserved organization of genes in trypanosomatids. *Molecular and Biochemical Parasitology* **94**, 249–264.

CANO, M. I., GRUBER, A., VAZQUEZ, M., CORTÉS, A., LEVIN, M. J., GONZALEZ, A, DEGRAVE, W., RONDINELLI, E., RAMIREZ, J. L., ALONSO, C., REQUENA, J. M. & DA SILVEIRA, F. J. (1995). Molecular karyotype of clone CL Brener chosen for the *Trypanosoma cruzi* genome project. *Molecular and Biochemical Parasitology* **71**, 273–278.

DEGRAVE, W., DE MIRANDA, A. B., AMORIM, A., BRANDAO, A., ASLETT, M. & VANDEYAR, M. (1997). TcruziDB, and integrated database, and the WWW information server for the *Trypanosoma cruzi* genome project. *Memorias do Instituto Oswaldo Cruz* **92**, 805–809.

FERRARI, I., LORENZI, H., SANTOS, M. R., BRANDARIZ, S., REQUENA, J. M., SCHIJMAN, A., VAZQUEZ, M., DA SILVEIRA, J. F., BEN-DOV, C., MEDRANO, C., GHIO, S., LOPEZ BERGAMI, P., CANO, I., ZINGALES, B., URMENYI, T. P., RONDINELLI, E., GONZALEZ, A., CORTES, A., LOPEZ, M. C., THOMAS, M. C., ALONSO, C., RAMIREZ, J. L., CHIURRILLO, M. A., ALDAO, R. R., LEVIN, M. J. *et al.* (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. *Memorias do Instituto Oswaldo Cruz* **92**, 843–852.

FROHME, M., HANKE, J., ASLUND, L., PETTERSSON, U. & HOHEISEL, J. D. (1998*a*). Selective generation of chromosomal cosmid libraries within the *Trypanosoma cruzi* genome project. *Electrophoresis* **19**, 478–481.

FROHME, M., HANKE, J., LAURANT, J. P., SWINDLE, J. & HOHEISEL, J. D. (1998*b*). Hybridization mapping of *Trypanosoma cruzi* chromosomes III and IV. *Electrophoresis* **19**, 482–485.

HENRIKSSON, J., ASLUND, L. & PETTERSSON, U. (1996). Karyotype variability in *Trypanosoma cruzi*. *Parasitology Today* **12**, 108–114.

HENRIKSSON, J., PORCEL, B., RYDAKER, M., RUIZ, A., SABAJ, V., GALANTI, N., CAZZULO, J. J., FRASCH, A. C. C. & PETTERSSON, U. (1995). Chromosome specific markers reveal conserved linkage groups in spite of extensive chromosomal size variation in *Trypanosoma cruzi*. *Molecular and Biochemical Parasitology* **73**, 63–74.

IVENS, A. C. & BLACKWELL, J. M. (1996). Unravelling the *Leishmania* genome. *Current Opinion in Genetics and Development* **6**, 704–710.

IVENS, A. C. & BLACKWELL, J. M. (1999). Progress in the *Leishmania* genome project. *Parasitology Today*. In Press).

IVENS, A., BAGHERZADEH, A., LEWIS, S. M., ZHANG, L., CHAN, H. M. & SMITH, D. F. (1998). A physical map of the *Leishmania major* genome. *Genome Research* **8**, 135–145.

MELVILLE, S. E. (1997). Parasite genome analysis. Genome research in *Trypanosoma brucei*: chromosome size polymorphism and its relevance to genome mapping and analysis. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **91**, 116–120.

MELVILLE, S. E. (1998). The African trypanosome genome project. Focus on the future. *Parasitology Today* **14**, 128–130.

MELVILLE, S. E., GERRARD, C. S. & BLACKWELL, J. M. (1999). Multiple causes of size variation in the diploid megabase chromosomes of African trypanosomes. *Chromosome Research* (In Press).

MELVILLE, S. E., LEECH, V., GERRARD, C. S., TAIT, A. & BLACKWELL, J. M. (1998*b*). The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* and the assignment of chromosome markers. *Molecular & Biochemical Parasitology* **94**, 155–173.

MELVILLE, S. E., MAJIWA, P. & DONELSON, J. (1998*a*). Resources available from the African trypanosome genome project. *Parasitology Today* **14**, 3–4.

MYLER, P. J., AUDELMAN, L., DEVOS, T., HIXSON, G., KISER, P., LEMLEY, C., MAGNESS, C., RICKEL, E., SISK, E., SUNKIN, S., SWARTZEL, S., WESTLAKE, T., BASTIEN, P., FU, G., IVENS, A. & STUART, K. (1999). *Leishmania major* Friedlin chromosome 1 has only two polycistronic units of protein encoding genes. *Proceedings of the National Academy of Sciences, USA*, (In Press).

SANTOS, M. R., CANO, M. I., SCHIJMAN, A., LORENZI, H., VAZQUEZ, M., LEVIN, M. J., RAMIREZ, J. L., BRANDAO, A., DEGRAVE, W. M. & DA SILVEIRA, J. F. (1997). The *Trypanosoma cruzi* genome project: nuclear karyotype and gene mapping of clone CL Brener. *Memorias do Instituto Oswaldo Cruz* **92**, 821–828.

TURNER, C. M. R., MELVILLE, S. E. & TAIT, A. (1997). A proposal for karyotype nomenclature in *Trypanosoma brucei*. *Parasitology Today* **13**, 5–6.

VERDUN, R. E., DI PAOLO, N., URMENYI, T. P., RONDINELLI, E., FRASCH, A. C. & SANCHEZ, D. O. (1988). Gene discovery through expressed sequence Tag sequencing in *Trypanosoma cruzi*. *Infection and Immunity* **66**, 5393–5398.

WINCKER, P., RAVEL, C., BLAINEAU, C., PAGES, M., JAUFFRET, Y., DEDET, J.-P. & BASTIEN, P. (1996). The *Leishmania* genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. *Nucleic Acids Research* **24**, 1688–1694.

ZINGALES, B., PEREIRA, M. E., ALMEIDA, K. A., UMEZAWA, E. S., NEHME, N. S., OLIVERIA, R. P., MACEDO, A. & SOUTO, R. P. (1997). Biological parameters and molecular markers of clone CL Brener – the reference organism of the *Trypanosoma cruzi* genome project. *Memorias do Instituto Oswaldo Cruz* **92**, 811–814.