Gene Action and Cellular Function in Parasitic Protozoa

Satellite Meeting of the 18th International Congress of Biochemistry and Molecular Biology Organized by K. Matthews and K. Gull (School of Biological Sciences, University of Manchester) and Edited by K. Matthews. Held at the Chancellor's Conference Centre, University of Manchester, 13–15 July 2000.

Genomic organization and gene function in *Leishmania*

P. J. Myler*[1], E. Sisk*, P. D. McDonagh*, S. Martinez-Calvillo*, A. Schnaufer*, S. M. Sunkin*, S. Yan*, R. Madhubala†, A. Ivens‡ and K. Stuart*

*Seattle Biomedical Research Institute, 4 Nickerson Street, Seattle, WA 98109-1651, U.S.A., †School of Life Sciences, Jawaharla Nehru University, New Delhi, 110067, India, and ‡Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, U.K.

Abstract

Sequencing of the *Leishmania major* Friedlin genome is well underway with chromosome 1 (Chr1) and Chr3 having been completely sequenced, and Chr4 virtually complete. Sequencing of several other chromosomes is in progress and the complete genome sequence may be available as soon as 2003. A large proportion ($\approx 70\%$) of the newly identified genes remains unclassified, with many of these being potentially *Leishmania*- (or kinetoplastid-) specific. Most interestingly, the genes are organized into large (> 100–300 kb) polycistronic clusters of adjacent genes on the same DNA strand. Chr1 contains two such clusters organized in a 'divergent' manner, i.e. the mRNAs for the two sets of genes are both transcribed towards the telomeres. Chr3 contains two 'convergent' clusters, with a single 'divergent' gene at one telomere, with the two large clusters separated by a tRNA gene. We have characterized several genes from the LD1 (*Leishmania* DNA 1) region of Chr35. *BT1* (formerly *ORFG*) encodes a biopterin transporter and *ORFF* encodes a nuclear protein of unknown function. Immunization of mice with recombinant antigens from these genes results in significant reduction in parasite burden following *Leishmania* challenge. Recombinant ORFF antigen shows promise as a serodiagnostic. We have also developed a tetracycline-regulated promoter system, which allows us to modulate gene expression in *Leishmania*.

Key words: genome sequencing, regulatable promoter, serodiagnosis, transcription, vaccine.

Abbreviations used: Chr, chromosome; LD1, *Leishmania* DNA 1; LmjF, *Leishmania major* Friedlin; $PrRNA_{tc}$, rRNA promoter.

[1]To whom correspondence should be addressed (e-mail mylerpj@sbri.org).

Introduction

The numerous human-infective *Leishmania* spp. cause a spectrum of diseases with pathologies ranging from asymptomatic to lethal, resulting in widespread human suffering and death, as well as substantial economic loss. There is a correlation between the parasite species and disease manifestation [1], but host factors play an important role in disease pathology. Diagnosis of *Leishmania* infection is insensitive, non-specific and labour-intensive; current chemotherapeutic agents are unsuitable because of their high toxicity, and there are no approved vaccines. Thus a greater knowledge of *Leishmania* biochemistry and genetics is sorely needed. The *Leishmania* genome size is ≈ 34 Mb and the chromosomes range in size from 0.3 to 2.8 Mb [2,3]. The *Leishmania* karyotype is conserved among *Leishmania* species (albeit with considerable size polymorphism) and the genes are syntenic [2,4,5], except that the Old World species have 36 chromosomes [2] and the New World species have 35 (*L. braziliensis* complex) or 34 (*L. mexicana* complex) [4].

Genome sequencing

In 1994, the *Leishmania* Genome Network (LGN) was set up under the auspices of the World Health Organization to initiate a *Leishmania* genome project, and *L. major* MHOM/IL/81/Friedlin (LmjF) was subsequently selected as the reference strain. A first-generation cosmid contig map of the

entire genome was constructed [6], and cosmid-based genomic sequencing began in 1996. The sequence of the smallest (285 kb) *Leishmania* chromosome (Chr1) was completed in 1998, and 79 protein-coding genes were identified [7]. Remarkably, these genes are organized into two large polycistronic units, with the first 29 genes on one DNA strand and the remaining 50 genes on the other stand, such that their mRNAs are transcribed in a 'divergent' manner towards the telomeres. The 257-kb 'informational' region (containing the protein-coding genes) is flanked by telomeric and sub-telomeric sequences that differ in size by $\approx$ 29 kb between Chr1 homologues [8]. The sequence of Chr3 (385 kb) has been completed recently in the U.S.A., as well as 39 cosmids representing $\approx$ 1.2 Mb of the 2.4-Mb Chr35, and two cosmids each from Chr22, Chr27 and Chr2. In Europe, Chr4 (408 kb), and most of Chr19 and Chr23 have been completed and sequencing has begun on Chr5, Chr13, Chr14 and Chr21. Currently, 3.3 Mb of completed sequence and 2.1 Mb of unfinished sequence has been submitted to the sequence databases. In addition, genome survey sequences (GSS) representing another 4 Mb have been submitted from Washington University (St. Louis, MO, U.S.A.). The ongoing sequencing progress can be monitored using the SBRI Leishgen website (http://204.203.14.2/lmjf) or the LGN website (http://www.ebi.ac.uk/parasites/leish.html).

Gene organization

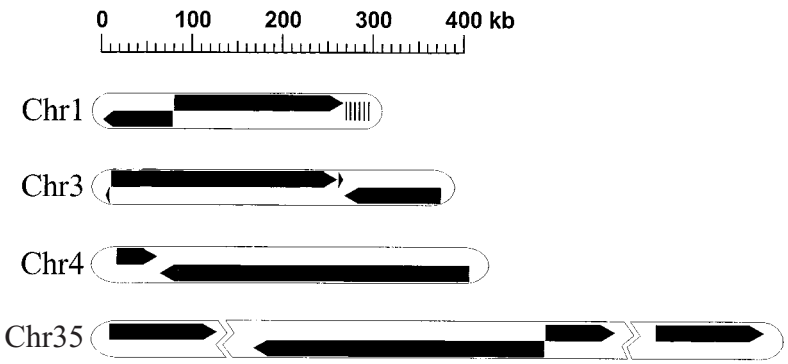
As was the case with Chr1, the genes on the other chromosomes are organized into large clusters with many genes adjacent on the same DNA strand (see Figure 1). Interestingly, the gene clusters can be either divergent (as on Chr1), where the mRNAs are transcribed toward the telomeres, or 'convergent' (on Chr3 and Chr4), where they are transcribed away from the telomeres. Indeed, it appears that these two organizations may occur on adjacent portions of the same chromosomes (e.g. Chr3). This gene organization is consistent with the previously observed polycistronic transcription of protein-coding genes in *Leishmania* (and other kinetoplastida) and subsequent processing to form mature mRNAs [9]. Recent nuclear run-on data for LmjF Chr1 supports this interpretation, since it appears that transcription of the protein-coding strand is significantly greater than that of the non-coding strand (S. Martinez-Calvillo and P. J. Myler, unpublished work). The processing of the precursor RNA, which occurs co-transcriptionally [10], involves co-ordinated 3′ polyadenylation of the upstream mRNA and addition of a 39-nt spliced leader sequence to the 5′ ends of the downstream mRNAs (*trans*-splicing). Polypyrimidine tracts within the intergenic regions provide the signals for this processing [11,12]. No introns have been discovered to date within any of the *Leishmania* protein-coding genes, although the recent discovery of *cis*-splicing in other trypanosomatids [13] suggests that this may not hold true for all *Leishmania* genes.

To date, almost 1000 complete genes have been identified from the cosmid sequencing, although this number is rapidly increasing and will no doubt be greater at the time of publication.

Figure 1

Gene organization of *Leishmania* chromosomes

The gene clusters on Chr1, Chr3, Chr4 and Chr35 are indicated by the thick lines, with the direction of mRNAs indicated by the arrowheads at the end of each line. The vertical lines at the right-hand end of Chr1 denote sub-telomeric repeat sequences. The gaps in Chr35 represent regions not yet sequenced.



The rRNA genes have been identified, but only one tRNA gene has been found. Interestingly, it is located at the junction of two convergent gene clusters on Chr3. The largest ($\approx 70\,\%$) gene category is that of the unclassified genes. Some of these represent genes encoding predicted proteins with sequence homology to proteins with unknown functions in other organisms, or those that contain uninformative sequence motifs, but most encode proteins with no identifying features or sequence similarities (other than to genes in other trypanosomatids). These may represent genes that have parasite-specific functions, or which are diverged sufficiently as to have no significant sequence similarity to their functional homologues in other species.

The gene distribution seen on Chr1, Chr3, Chr4 and Chr35 indicates that *Leishmania* genes do not tend to cluster into prokaryote-like operons of genes with similar function. Interestingly, however, some regions appear to have a higher-than-expected concentration of large genes with no similarity to those in other organisms. The gene density (1 gene/3.7 kb) observed within the informational regions of Chr1, Chr3 and Chr4 extrapolates to a total of ≈ 8600 protein-coding genes in the entire *Leishmania* genome. Given that $\approx 70\,\%$ of the protein-coding genes have no currently identified function, it is reasonable to infer that completion of the *Leishmania* genome sequence will identify 4000–5000 genes with potentially parasite-specific functions. Statistical analyses of the nucleotide content of Chr1 reveals a striking, non-random purine bias and GC skew that is correlated with the two polycistronic units of protein-coding genes [14]. These findings suggest that novel transcription processes in *Leishmania* may be responsible for the nucleotide bias, which in turn affects chromosomal gene organization. They also suggest that the junction region between the two divergent polycistronic gene clusters on Chr1 may be a candidate for an origin of DNA replication.

Leishmania DNA I (LD1) genes

The LD1 region, which is located ≈ 100 kb from one telomere of Chr35, is amplified in $\approx 15\,\%$ of *Leishmania* strains examined [15]. Comparison of a 131-kb LmjF sequence with the 35-kb sequences obtained from *L. donovani* and *L. infantum* [16–18] shows considerably more sequence conservation (91–96 %) within the protein-coding open reading frames than within the non-coding regions (79–85 %). Using antibodies against recombinant protein (rORFF), we demonstrated that ORFF is localized to the parasite nucleus [19]. Purified rORFF protein was found to be a more sensitive serodiagnostic than the total soluble antigen now in common use, and appeared specific for the *L. donovani* complex [20]. Immunization of mice with rORFF and rORFG proteins, individually and in combination, showed that they stimulated both humoral and cellular responses, and provided a significant degree of protection against subsequent challenge with *L. donovani*, with up to a 60 % reduction in liver and spleen amastigote burdens [21].

Using successive rounds of gene replacement of the three *ORFG* genes in *L. donovani* LSB-51.1 [22], an *ORFG*-null mutant was obtained. The null mutants were unable to survive in standard growth medium without added biopterin, and showed essentially no uptake of radioactive biopterin. Thus *ORFG* encodes a biopterin transporter and has been renamed *BT1* [23]. These results suggest that uptake of biopterin by BT1 is critical for cell survival at physiological concentrations of biopterin. The *BT1*-null mutants showed significantly slower growth as promastigotes, even in the presence of supplemental biopterin and folate, suggesting that their ability to grow in this medium is due to uptake of biopterin via passive diffusion or a secondary (and lower-affinity) transporter. Conversely, amplification and overexpression of *BT1* confers a significant growth advantage in both naturally isolated and recombinant cell lines, perhaps providing the selective pressure for the frequent amplification of LD1 seen in *Leishmania* isolates [24,25]. This advantage appears to extend to macrophage infectivity, since the naturally occurring isolates with amplified *BT1* show significantly higher infectivity of macrophages *in vitro*, while the null mutant showed a significant reduction in infectivity.

Development of a regulatable promoter system in *Leishmania*

The transcription-initiation site of *L. donovani* rRNA genes was mapped 1020 bp upstream of the 18 S rRNA gene and the promoter was functionally characterized using transient-transfection studies [26]. Three domains (-76 to -57, -46 to -27 and -6 to $+4$ relative to the transcription-initiation site) were found to mediate promoter activity, suggesting that the rRNA is not dissimilar to that of other eukaryotes. Similar results were obtained with the *L. major* rRNA promoter, which

was active in *L. major*, *L. donovani*, *L. aethiopica* and *L. infantum*, while the *L. donovani* promoter showed reduced activity in *L. major*.

In order to adapt the prokaryotic tetracycline-responsive repressor/operator system to *L. donovani*, promastigotes expressing the tetracycline-repressor gene in the α-tubulin locus were transfected with a construct containing a phleomycin-resistance/luciferase fusion gene driven by the rRNA promoter ($PrRNA_{tc}$) containing two copies of the tetracycline operator sequence. The construct was targeted into the rRNA non-transcribed intergenic region in the reverse orientation relative to transcription of the *rRNA* gene. This system showed an increase in luciferase expression by more than 200-fold in the presence of tetracycline. Comparison with similar constructs (but lacking $PrRNA_{tc}$) integrated into the α-tubulin and rRNA loci showed that luciferase expression driven by the down-regulated $PrRNA_{tc}$ (i.e. in the absence of tetracycline) was ≈ 100-fold lower than expression by pol II (tubulin locus) and ≈ 5000-fold lower than expression by pol I (rRNA locus). Conversely, luciferase activity driven by the up-regulated $PrRNA_{tc}$ (i.e. in the presence of tetracycline) was ≈ 2–3-fold higher than that from pol II expression at the tubulin locus.

However, in these first-generation constructs, even in the absence of tetracycline, the luciferase activity was ≈ 100-fold above background, so a second-generation construct was developed, in which the selectable marker (hygromycin resistance) was expressed from an endogenous promoter on the opposite strand to the reporter gene (luciferase) under control of the tetracycline-regulated promoter. When targeted into either the rRNA (Chr27) or LD1 (Chr35) loci, this construct showed regulation of luciferase expression by over three orders of magnitude. Clones with the construct in the rRNA locus showed 10-fold higher luciferase activities than those in the LD1 locus, both in the presence and absence of tetracycline. Integration into the LD1 locus showed luciferase levels close to background in the absence of tetracycline. Induction of luciferase expression was rapid, becoming elevated by up to two orders of magnitude within 5 h of tetracycline addition, with full induction by 24 h. In contrast, luciferase activity decreased 3–4-fold within 4 h of tetracycline removal, then slowly declined until it reached background levels after about 1 week. The level of luciferase activity was dependent on tetracycline concentration, with a concentration of 0.1 μg/ml being sufficient for maximal induction, while 0.001 μg/ml resulted in little or no induction. This technology is currently being transferred into LmjF, the *Leishmania* strain being sequenced. The ability to tightly regulate promoter activity will be extremely useful for studying gene function in *Leishmania* and several applications of this inducible system are in progress.

Conclusions and future directions

The pace of sequence generation in the *Leishmania* genome project has increased substantially over the past year and it is likely that the entire genome will be complete by 2003. Thus, the next 2 years will see a massive explosion in the number of *Leishmania* gene sequences available for study. Already, serious efforts are underway in several laboratories to implement the next stages of high-throughput genome-wide analyses, such as DNA microarrays [27] and proteomics [28]. When combined with new molecular tools (such as the regulatable promoter system) for analyses of *Leishmania* biology, these studies will probably cause a paradigm shift in our quest to understand and control this parasite.

We thank all the members, past and present, of the SBRI, Sanger Centre and EuLeish consortium *Leishmania* sequencing teams for their tireless efforts in sequencing the LmjF genome, as well as other members of the Myler, Stuart and Madhubala laboratories for their work on *Leishmania* gene function. We also thank Stephen Beverley, Paul Englund, Angela Cruz, and members of their laboratories for sharing data before publication. This work was funded by the National Institute of Allergy and Infectious Diseases (NIAID) and the Burroughs-Wellcome Fund in the U.S.A., as well as the European Commission and Beowulf Genomics (Wellcome Trust) in Europe.

References

1 Shaw, J. J. and Lainson, R. (1987) in The Leishmaniases in Biology and Medicine, Ecology and Epidemiology: New World (Peters, W. and Killick-Kendrick, R., eds), pp. 291–361, Academic Press, London
2 Wincker, P., Ravel, C., Blaineau, C., Pages, M., Jauffret, Y., Dedet, J., Bastien, P. and Dedet, J. P. (1996) Nucleic Acids Res. **24**, 1688–1694
3 Bastien, P., Blaineau, C., Britto, C., Dedet, J.-P., Dubessay, P., Pagès, M., Ravel, C., Winker, P., Blackwell, J. M., Leech, V. et al. (1998) Parasitol. Today **14**, 301–303
4 Britto, C., Ravel, C., Bastien, P., Blaineau, C., Pagès, M., Dedet, J. P. and Wincker, P. (1998) Gene **222**, 107–117
5 Ravel, C., Dubessay, P., Britto, C., Blaineau, C., Bastien, P. and Pagès, M. (1999) Nucleic Acids Res. **27**, 2473–2477
6 Ivens, A. C., Lewis, S. M., Bagherzadeh, A., Zhang, L., Chang, H. M. and Smith, D. F. (1998) Genome Res. **8**, 135–145
7 Myler, P. J., Audleman, L., deVos, T., Hixson, G., Kiser, P., Lemley, C., Magness, C., Rickell, E., Sisk, E., Sunkin, S. et al. (1999) Proc. Natl. Acad. Sci. U.S.A. **96**, 2902–2906

8 Sunkin, S. M., Kiser, P., Myler, P. J. and Stuart, K. D. (2000) Mol. Biochem. Parasitol. **109**, 1–15
9 Perry, K. and Agabian, N. (1991) Experientia **47**, 118–128
10 Ullu, E., Matthews, K. R. and Tschudi, C. (1993) Mol. Cell. Biol. **13**, 720–725
11 Matthews, K. R., Tschudi, C. and Ullu, E. (1994) Genes Dev. **8**, 491–501
12 Lopez-Estrano, C., Tschudi, C. and Ullu, E. (1998) Mol. Cell. Biol. **18**, 4620–4628
13 Mair, G., Shi, H., Li, H., Djikeng, A., Aviles, H. O., Bishop, J. R., Falcone, F. H., Gavrilescu, C., Montgomery, J. L., Santori, M. I. et al. (2000) RNA **6**, 163–169
14 McDonagh, P. D., Myler, P. J. and Stuart, K. D. (2000) Nucleic Acids Res. **28**, 2800–2803
15 Tripp, C. A., Myler, P. J. and Stuart, K. (1991) Mol. Biochem. Parasitol. **47**, 151–160
16 Myler, P. J., Tripp, C. A., Thomas, L., Venkataraman, G. M., Merlin, G. and Stuart, K. D. (1993) Mol. Biochem. Parasitol. **62**, 147–152
17 Myler, P. J., Lodes, M. J., Merlin, G., deVos, T. and Stuart, K. D. (1994) Mol. Biochem. Parasitol. **66**, 11–20
18 Myler, P. J., Venkataraman, G. M., Lodes, M. J. and Stuart, K. D. (1994) Gene **148**, 187–193
19 Ghosh, A., Raj, V. S., Madhubala, R., Myler, P. J. and Stuart, K. D. (1999) Exp. Parasitol. **93**, 225–230
20 Raj, V. S., Ghosh, A., Dole, V., Madhubala, R., Myler, P. J. and Stuart, K. D. (1999) Am. J. Trop. Med. Hyg. **61**, 482–487
21 Dole, V., Raj, V S., Ghosh, A., Madhubala, R., Myler, P. J. and Stuart, K. D. (2000) Vaccine, in the press
22 Lodes, M. J., Merlin, G., deVos, T., Ghosh, A., Madhubala, R., Myler, P. J. and Stuart, K. (1995) Mol. Cell. Biol. **15**, 6845–6853
23 Lemley, C., Yan, S., Cunningham, M. W., Dole, V., Raj, V. S., Ghosh, A., Madhubala, R., Beverley, S. M., Myler, P. J. and Stuart, K. D. (1999) Mol. Biochem. Parasitol. **104**, 93–105
24 Stuart, K. D. (1991) Parasitol. Today **7**, 158–159
25 Segovia, M. (1994) Ann. Trop. Med. Parasitol. **88**, 123–130
26 Yan, S., Lodes, M. J., Fox, M., Myler, P. J. and Stuart, K. (1999) Mol. Biochem. Parasitol. **103**, 197–210
27 DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. and Trent, J. M. (1996) Nat. Genet. **14**, 457–460
28 Humphery-Smith, I., Cordwell, S. J. and Blackstock, W. P. (1997) Electrophoresis **18**, 1217–1242

Received 30 June 2000

Life-cycle differentiation in *Trypanosoma brucei*: molecules and mutants

E. Hendriks, F. J. van Deursen, J. Wilson, M. Sarkar, M. Timms and K. R. Matthews[1]
2.205 Stopford Building, School of Biological Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, U.K.

Abstract

Differentiation between bloodstream and tsetse midgut procyclic forms during the life cycle of the African trypanosome is an attractive model for the analysis of stage-regulated events. In particular, this transformation occurs synchronously, there are well-defined markers for stage-regulated processes and cell lines with specific defects in differentiation have been identified. This combination of tools, combined with the developing *Trypanosoma brucei* genome database is allowing its underlying controls to be investigated at the molecular and cytological levels. This paper examines some recent discoveries that illuminate some of the key events during trypanosome life-cycle progression.

Overview

African trypanosomiasis is spread between mammalian hosts in sub-Saharan Africa by the blood-feeding tsetse fly. The parasite undergoes many changes as it is transmitted between the blood of a mammalian host and the gut of the tsetse fly, these being required for adaptation to the very different environmental conditions encountered [1]. In particular, the parasite undergoes metabolic development in order to successfully move from the glucose-rich bloodstream to the tsetse midgut, where proline is the main energy source [2]. The trypanosome also changes the major proteins that comprise its surface coat [3,4] and undergoes morphological development from the bloodstream trypomastigote to the tsetse procyclic stage [5]. These changes to the parasite's biochemistry, morphology and profile of expressed genes and proteins are highly regulated and provide a tractable system for the analysis of life-cycle differentiation in these evolutionarily ancient protozoan parasites.

Key words: cell cycle, genome, trypanosome.

Abbreviations used: VSG coat, variant-specific surface glycoprotein coat; GPI-PLC, glycophosphatidylinositol phospholipase C; DiD-1, Defective in Differentiation clone 1.

[1]To whom correspondence should be addressed (e-mail keith.matthews@man.ac.uk).

Bloodstream-to-procyclic-form differentiation provides a model for the analysis of regulated events in the trypanosome life cycle

Although trypanosomes grow to high levels of parasitaemia in infected mammalian hosts, only