Med Microbiol Immunol (2001) 190: 9–12
DOI 10.1007/s004300100070

**Peter J. Myler · Stephen M. Beverley · Angela K. Cruz
Deborah E. Dobson · Alasdair C. Ivens
Paul D. McDonagh · Rentala Madhubala
Santiago Martinez-Calvillo · Jeronimo C. Ruiz
Alka Saxena · Ellen Sisk · Susan M. Sunkin
Elizabeth Worthey · Shaofeng Yan · Kenneth D. Stuart**

The *Leishmania* genome project: new insights into gene organization and function

Published online: 15 August 2001

Abstract The sequencing of *Leishmania major* Friedlin chromosome 1 (Chr1), Chr3, and Chr4 has been completed, and several other chromosomes are well underway. The complete genome sequence should be available by 2003. Over 1,000 full-length new genes have been identified, with the majority (~75%) having unknown function. Many of these may be *Leishmania* (or kinetoplastid) specific. Most interestingly, the genes are organized into large (>100–500 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. Nuclear run-on analysis suggests that transcription is initiated in both directions within the "divergent" region. Chr3 and Chr4 contain two "convergent" clusters, with a single "divergent" gene at one telomere of Chr3. Sequence analysis of several genes from the LD1 region of Chr35 indicates a high degree of sequence conservation between *L. major* and *L. donovani*/*L. infantum* within protein-coding open reading frames (ORFs), with a lower degree of conservation within the non-coding regions. Immunization of mice with recombinant antigen from two of these genes, *BT1* (formerly *ORFG*) and *ORFF*, 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*.

P.J. Myler (✉) · P.D. McDonagh · S. Martinez-Calvillo
A. Saxena · E. Sisk · S.M. Sunkin · E. Worthey
S. Yan · K.D. Stuart
Seattle Biomedical Research Institute,
4 Nickerson Street, Seattle, WA 98109-1651, USA
E-mail: mylerpj@sbri.org
Tel.: +1-206-2848846
Fax: +1-206-2840313

S.M. Beverley · D.E. Dobson
Department of Molecular Microbiology,
Washington University School of Medicine,
660S. Euclid Ave., St. Louis, MO 63110-1093, USA

A.K. Cruz · J.C. Ruiz
Departamento de Bioquimica,
Faculdade de Medicina de Ribeirao Preto,
Universidade de Sao Paulo, Sao Paulo, Brazil

A.C. Ivens
Sanger Centre, Wellcome Trust Genome Campus,
Hinxton, Cambridge, CB10 1SA, UK

R. Madhubala
School of Life Sciences,
Jawaharla Nehru University, New Delhi, 110067, India

Genome sequencing

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. The *Leishmania* genome size is ~34 Mb and the chromosomes range in size from 0.3 to 2.8 Mb [1, 23]. The *Leishmania* karyotype is conserved among *Leishmania* species (albeit with considerable size polymorphism) and the genes are syntenic [2, 20, 23], except that the Old Word species have 36 chromosomes [23] and the New World species have 35 (*L. braziliensis* complex) or 34 (*L. mexicana* complex) [2]. The *Leishmania* Genome Network (LGN) was set up to initiate a *Leishmania* genome project using *L. major* MHOM/IL/81/Friedlin (LmjF) as the reference strain. A first generation cosmid contig map of the entire genome was constructed [7], and cosmid-based genomic sequencing began at Seattle Biomedical Research Institute in the USA, the Sanger Centre in the UK and several European laboratories as part of the EuLeish consortium. By the end of 2000, almost 10 Mb of cosmid and PAC sequence has been

generated, representing more than 25% of the LmjF genome. In addition, Genome Survey Sequences (GSS) from random genomic clones, as well as cosmid and BAC end-sequences provide another 5 Mb of single pass sequence. Several chromosomes have been completely sequenced, revealing a very interesting pattern of gene organization. The sequence of smallest *Leishmania* chromosome (Chr1) was completed in 1998 [17], with Chr3 and Chr4 being completed within the last year. A large number of cosmids and PACs from several other chromosomes (Chr2, Chr5, Chr13, Chr14, Chr19, Chr21, Chr23 and Chr35) have been sequenced and sequencing of random shotgun clones from several other chromosomes has begun. Ongoing progress of the LmjF sequencing project can be followed at http://www.ebi.ac.uk/parasites/leish.html, http://www.genome.sbri.org/lmjf, http://www.sanger.ac.uk/Projects/L_major and http://paradb.cis.upenn.edu/leish/index.html.

At the time of writing, over 1,000 complete and 2,000 incomplete new protein-coding genes have been identified, along with several rRNA, spliced leader (SL) RNA and tRNA genes. The density of protein-coding genes appears to be relative constant (one gene per 3.7 kb and 55% coding sequence) throughout the regions sequenced to date and predicts a total of ~8,600 genes for the entire LmjF genome. The genes do not cluster into prokaryote-like operons of genes with similar function, but some regions appear to have a higher-than-expected concentration of large genes with no similarity to those in other organisms. A significant proportion (5–10%) of the genes occur in more than one copy, often as tandem duplications, but sometimes as larger interspersed or scattered gene families. The sequence differences between members of these gene families is often substantial, suggesting an ancient divergence, perhaps to carry out related, but distinct, functions. Categorization of the complete genes into 13 groups according to their predicted function (Table 1) reveals that the vast majority (~75%) remain unclassified. Some of these represent genes encoding predicted proteins with sequence similarity to proteins of unknown functions in other organisms, or that contain relatively uninformative sequence motifs, but many 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 sufficiently diverged as to have no significant sequence similarity to their functional homologs in other species.

Table 1 Categorization of the complete genes according to their predicted function

Category	SBRI	Sanger	EuLeish	Total	%
Metabolism	24	6	10	40	4
Energy generation	7	1	10	18	2
Cell growth, division and DNA synthesis	5	0	1	6	1
Transcription and RNA processing	8	0	2	10	1
Protein synthesis	14	4	11	29	3
Protein destination	23	6	10	39	4
Transport	15	2	13	30	3
Intracellular trafficking	5	0	7	12	1
Signal transduction	6	2	11	19	2
Cellular organization/biogenesis	8	2	13	23	2
Cell rescue, defense and aging	16	0	13	29	3
Structural RNA	13	0	0	13	1
Unclassified	251	108	427	786	75
Total	395	131	528	1,054	100

Chromosomal organization

The 285-kb Chr1 contains a 257-kb "informational" region containing the 79 protein-coding genes [17], flanked by telomeric and sub-telomeric sequences that differ in size by ~29 kb between Chr1 homologues [21]. Remarkably, the 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 385-kb Chr3 contains only small telomeric regions and the 94 protein-coding genes are organized into two large "convergent" polycistronic units of 64 and 29 genes that are transcribed away from the telomeres, and a single gene at the "left" telomere transcribed toward the telomere in a "divergent" manner with the larger of these two units. The 115 protein-coding genes on Chr4 (408 kb) are organized into two "convergent" polycistronic units of 28 and 87 genes. The genes on other chromosomes also appear to be organized into similar large "divergent" and "convergent" polycistronic clusters (Fig. 1). This gene organization is consistent with the previously observed polycistronic transcription of protein-coding

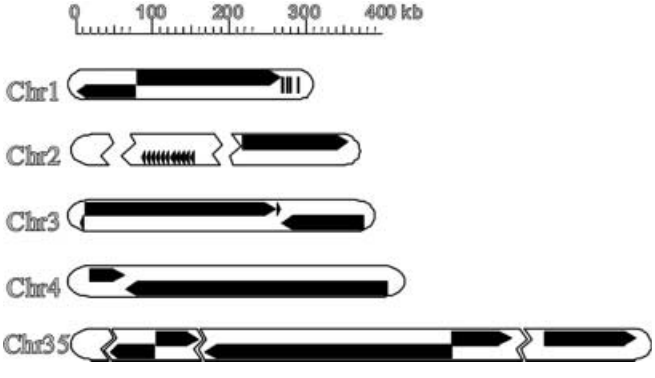

Fig. 1 Gene organization of *Leishmania* chromosomes. The gene clusters on Chr1, Chr3, Chr4 and Chr35 are indicated by the *thick lines*, with the direction of mRNAs indicated by the *arrowheads* at the end of each line. The *vertical lines* at the "*right*" end of Chr1 denote sub-telomeric repeat sequences. The *arrowheads* in Chr2 represent SL gene clusters, while the *single arrowhead* between the gene clusters on Chr3 represents a tRNA. The *gaps* in Chr2 and Chr35 represent regions not yet sequenced, and the relative orientation of the gene clusters within the internal segments is yet to be determined

genes in *Leishmania* (and other kinetoplastida) and subsequent processing to form mature mRNAs [18]. Interestingly, in more than one case (e.g., Chr3), one or more tRNA genes have been found between "convergent" units of protein-coding genes. A similar juxtaposition of RNA polymerase (Pol II) and Pol III transcriptional units has recently been observed in *Trypanosoma brucei* [11]. Nuclear run-on analyses indicate that transcription of the protein-coding strand of Chr1 is significantly greater than the non-coding strand, and the latter may not be transcribed at all. Experiments using UV irradiation of the cells before run-on, as well as RNase protection and primer extension analyses, suggest that transcription is initiated within a small (<200 bp) region within the region between the "divergent" polycistronic units of Chr1. Thus, it appears that this region contains a Pol II promoter, which has proven to be an elusive target in trypanosomatids to date. 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 [13]. An explanation for this finding is still lacking, but is likely to involve the polypyrimidine tracts within the intergenic regions that provide the signals for the co-transcriptional processing of the precursor RNA, which involves coordinated 3' polyadenylation of the upstream mRNA and the *trans*-splicing of a 39-nt SL (mini-exon) sequence to the 5' end of the downstream mRNA [9, 12, 22]. 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 [10], suggests that this may not hold true for all *Leishmania* genes.

Functional genomics

Comparison of the LD1 region near one telomere of Chr35 shows considerably more sequence conservation between *L. major* and *L. infantum*/*L. donovani* [14, 15, 16] within the protein-coding open reading frames (ORFs) (91–96%) than within the non-coding regions (79–85%). Functional characterization of two genes from the LD1 region indicated that *ORFG* encodes a biopterin transporter (BT1) [8], and the *ORFF* gene product is localized within the parasite nucleus [5]. A serodiagnostic assay developed using purified recombinant ORFF (rORFF) protein was more sensitive than the total soluble antigen now in common use, and appeared specific for the *L. donovani* complex [19]. Immunization of mice with rORFF and rORFG proteins, alone, and in combination, provided a significant degree of protection against subsequent challenge with *L. donovani* [4].

We have adapted the prokaryotic tetracycline-responsive repressor-operator system for studying the function of essential genes and expression of toxic gene products in *Leishmania*. The first generation system in *L. donovani* targeted a phleomycin resistance and luciferase fusion gene (*BLE-LUC*) into the rRNA non-transcribed intergenic region, in the reverse orientation relative to transcription of the *rRNA* gene, under the control of an rRNA promoter containing two copies of the tetracycline operator ($TETO_2$) sequence. This system showed background expression levels (i.e., in the absence of tetracycline) $\sim$100-fold lower than expression by Pol II (tubulin locus) and $\sim$5,000-fold lower than expression by Pol I (*rRNA* locus) and an increase of *BLE-LUC* expression by more than 200-fold in the presence of tetracycline [24]. A second generation of constructs, featuring marker (*HYG*) and reporter (*LUC*) genes in a "back-to-back" orientation (i.e., pointing away from each other on opposite DNA strands) showed very low background expression and three orders of magnitude increase on addition of tetracycline. The kinetics of induction were relatively rapid ($\sim$24 h), but down-regulation was slower ($\sim$200 h), and the level of expression was dependent on the tetracycline concentration.

Future perspective

Within the next 2 years, there will be 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 stage of high-throughput genome-wide analyses, such as DNA microarrays [3] and proteomics [6]. When combined with new molecular tools (such as the regulatable promoter system) for analyses of *Leishmania* biology, these studies will likely cause a paradigm shift in our quest to understand and control this parasite.

Acknowledgements We would like to thank all the members, past and present, of the *Leishmania* sequencing teams at SBRI, Sanger Centre, and EuLeish consortium for their tireless efforts in sequencing the LmjF genome, as well as other members of the Myler, Stuart, Madhubala, Beverley, and Cruz laboratories for their work on mapping and gene function. This work was funded by the National Institute of Allergy and Infectious Diseases (NIAID) and the Burroughs-Wellcome Fund in the USA, as well as the European Commission and Beowulf Genomics (Wellcome Trust) in Europe.

References

1. Bastien P, Blaineau C, Britto C, Dedet J-P, Dubessay P, Pagès M, Ravel C, Winker P, Blackwell JM, Leech V, Levick M, Norrish A, Ivens A, Lewis S, Bagherzadeh A, Smith D, Myler P, Stuart K, Cruz A, Ruiz JC, Schneider H, Sampaio I, Almeida R, Papadopoulou B, Shapira M, Belli S, Fasel N (1998) The complete chromosomal organization of the reference strain of the *Leishmania* genome project, *L. major* 'Friedlin'. Parasitol Today 14:301–303
2. Britto C, Ravel C, Bastien P, Blaineau C, Pagès M, Dedet JP, Wincker P (1998) Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World *Leishmania* genomes. Gene 222:107–117
3. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 14:457–460

4. Dole V, Raj VS, Ghosh A, Madhubala R, Myler PJ, Stuart KD (2001) Immunization with recombinant LD1 antigens protects against experimental leishmaniasis. Vaccine 19:423–430
5. Ghosh A, Raj VS, Madhubala R, Myler PJ, Stuart KD (1999) *Leishmania donovani*: characterization and expression of *ORFF*, a gene amplified from the LDI locus. Exp Parasitol 93:225–230
6. Humphery-Smith I, Cordwell SJ, Blackstock WP (1997) Proteome research: complementarity and limitations with respect to the RNA and DNA worlds. Electrophoresis 18:1217–1242
7. Ivens AC, Lewis SM, Bagherzadeh A, Zhang L, Chang HM, Smith DF (1998) A physical map of the *Leishmania major* Friedlin genome. Genome Res 8:135–145
8. Lemley C, Yan S, Cunningham MW, Dole V, Raj VS, Ghosh A, Madhubala R, Beverley SM, Myler PJ, Stuart KD (1999) The *Leishmania donovani* LD1 locus gene *ORFG* encodes a biopterin transporter (BT1). Mol Biochem Parasitol 104:93–105
9. Lopez-Estrano C, Tschudi C, Ullu E (1998) Exonic sequences in the 5' untranslated region of alpha-tubulin mRNA modulate trans splicing in *Trypanosoma brucei*. Mol Cell Biol 18:4620–4628
10. Mair G, Shi H, Li H, Djikeng A, Aviles HO, Bishop JR, Falcone FH, Gavrilescu C, Montgomery JL, Santori MI, Stern LS, Wang Z, Ullu E, Tschudi C (2000) A new twist in trypanosome RNA metabolism: *cis*-splicing of pre-mRNA. RNA 6:163–169
11. Marchetti MA, Tschudi C, Silva E, Ullu E (1998) Physical and transcriptional analysis of the *Trypanosoma brucei* genome reveals a typical eukaryotic arrangement with close interspersion of RNA polymerase II- and III-transcribed genes. Nucleic Acids Res 26:3591–3598
12. Matthews KR, Tschudi C, Ullu E (1994) A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. Genes Dev 8:491–501
13. McDonagh PD, Myler PJ, Stuart KD (2000) The unusual gene organization of *Leishmania* major chromosome 1 may reflect novel transcription processes. Nucleic Acids Res 28:2800–2803
14. Myler PJ, Tripp CA, Thomas L, Venkataraman GM, Merlin G, Stuart KD (1993) The LD1 amplified element from *Leishmania infantum* encodes a homolog of ribosomal protein L37. Mol Biochem Parasitol 62:147–152
15. Myler PJ, Lodes MJ, Merlin G, deVos T, Stuart KD (1994) An amplified DNA element in *Leishmania* encodes potential integral membrane and nucleotide binding proteins. Mol Biochem Parasitol 66:11–20
16. Myler PJ, Venkataraman GM, Lodes MJ, Stuart KD (1994) A frequently amplified region in *Leishmania* contains a gene that is conserved in prokaryotes and eukaryotes. Gene 148:187–193
17. Myler PJ, Audleman L, deVos T, Hixson G, Kiser P, Lemley C, Magness C, Rickell E, Sisk E, Sunkin S, Swartzell S, Westlake T, Bastien P, Fu G, Ivens A, Stuart K (1999) *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. Proc Natl Acad Sci USA 96:2902–2906
18. Perry K, Agabian N (1991) mRNA processing in the Trypanosomatidae. Experientia 47:118–128
19. Raj VS, Ghosh A, Dole V, Madhubala R, Myler PJ, Stuart KD (1999) Serodiagnosis of leishmaniasis with recombinant ORFF antigen. Am J Trop Med Hyg 61:482–487
20. Ravel C, Dubessay P, Britto C, Blaineau C, Bastien P, Pagès M (1999) High conservation of the fine-scale organisation of chromosome 5 between two pathogenic *Leishmania* species. Nucleic Acids Res 27:2473–2477
21. Sunkin SM, Kiser P, Myler PJ, Stuart KD (2000) The size difference between *Leishmania major* Friedlin chromosome one homologues is localized to sub-telomeric repeats at one chromosomal end. Mol Biochem Parasitol 109:1–15
22. Ullu E, Matthews KR, Tschudi C (1993) Temporal order of RNA-processing reactions in trypanosomes: rapid *trans* splicing precedes polyadenylation of newly synthesized tubulin transcripts. Mol Cell Biol 13:720–725
23. Wincker P, Ravel C, Blaineau C, Pages M, Jauffret Y, Dedet J, Bastien P, Dedet JP (1996) The *Leishmania* genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. Nucleic Acids Res 24:1688–1694
24. Yan S, Myler PJ, Stuart KD (2001) Tetracycline regulated gene expression in *Leishmania donovani*. Mol Biochem Parasitol 112:61–69