Leishmania major chromosome 3 contains two long convergent polycistronic gene clusters separated by a tRNA gene

E. A. Worthey¹, Santiago Martinez-Calvillo¹, Achim Schnaufer¹, Gautam Aggarwal¹, Jason Cawthra¹, Gholam Fazelinia¹, Chris Fong¹, Guoliang Fu², Melissa Hassebrock¹, Greg Hixson¹, Alasdair C. Ivens³, Patti Kiser¹, Felicia Marsolini¹, Erica Rickell¹, Reza Salavati¹, Ellen Sisk¹, Susan M. Sunkin¹, Kenneth D. Stuart¹,⁴ and Peter J. Myler¹,⁴,⁵,⁎

¹Seattle Biomedical Research Institute, 4 Nickerson Street, Seattle, WA 98109-1651, USA, ²Department of Pathology, University of Cambridge, Cambridge CB2 1QP, UK, ³Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK, ⁴Department of Pathobiology, University of Washington, Seattle, WA 98195, USA and ⁵Department of Medical Education and Biomedical Informatics, University of Washington, Seattle, WA 98195, USA

Received February 11, 2003; Revised April 17, 2003; Accepted May 14, 2003 DDBJ/EMBL/GenBank accession no. AC125735

ABSTRACT

Leishmania parasites (order Kinetoplastida, family Trypanosomatidae) cause a spectrum of human diseases ranging from asymptomatic to lethal. The ~33.6 Mb genome is distributed among 36 chromosome pairs that range in size from ~0.3 to 2.8 Mb. The complete nucleotide sequence of Leishmania major Friedlin chromosome 1 revealed 79 protein-coding genes organized into two divergent polycistronic gene clusters with the mRNAs transcribed towards the telomeres. We report here the complete nucleotide sequence of chromosome 3 (384 518 bp) and an analysis revealing 95 putative protein-coding ORFs. The ORFs are primarily organized into two large convergent polycistronic gene clusters with the mRNAs transcribed towards the telomeres. We report here the complete nucleotide sequence of chromosome 3 (384 518 bp) and an analysis revealing 95 putative protein-coding ORFs. The ORFs are primarily organized into two large convergent polycistronic gene clusters with the mRNAs transcribed towards the telomeres. In addition, a single gene at the left end is transcribed divergently towards the telomere, and a tRNA gene separates the two convergent gene clusters. Numerous genes have been identified, including those for metabolic enzymes, kinases, transporters, ribosomal proteins, spliceosome components, helicases, an RNA-binding protein and a DNA primase subunit.

INTRODUCTION

The Kinetoplastida are flagellated, primarily parasitic, protozoans found in terrestrial and aquatic environments. Kinetoplastid species cause disease in metazoa ranging from plants to vertebrates, resulting in extensive human suffering and death, as well as considerable economic loss from infection of livestock, wildlife and crops (1,2). Study of these organisms has been valuable for the investigation of fundamental molecular and cellular phenomena, such as RNA editing (3), mRNA trans-splicing (4), GPI anchoring of proteins (5), antigenic variation (6) and telomere organization (7). Their early evolutionary divergence makes comparison of their DNA and protein sequence with those of other eukaryotes, as well as prokaryotes, useful for the identification of ancient conserved motifs and their protein sequences may be a useful source of diversity for protein engineering.

Leishmania parasites belong to the family Trypanosomatidae and the numerous human-infective Leishmania spp. are responsible for a wide spectrum of human diseases with pathologies that range from asymptomatic to lethal (8). The World Health Organization (WHO) has estimated that there are over two million new cases of leishmaniasis each year in 88 countries, with 367 million people at risk (http://www.who.int/health-topics/leishmaniasis.htm). Whilst a correlation exists between disease manifestation and parasite species (8), host factors also play an important role in disease pathology (9). Leishmaniasis/HIV co-infection remains a significant problem (10). Diagnosis of Leishmania infection remains problematic; existing methods are labor-intensive and have sensitivity and specificity issues; current chemotherapeutic agents are unsuitable because of their high toxicity; and there are no approved vaccines (11,12). For these reasons, a greater knowledge of Leishmania biochemistry and genetics is sorely required.

The Leishmania karyotype is conserved among species (albeit with considerable size polymorphism among species), except that the Old World species (including Leishmania major) have 36 chromosomes (13) and the New World species are reported to have 35 (Leishmania braziliensis complex) or 34 (Leishmania mexicana complex) (14). More modest chromosome size polymorphisms exist within species amongst the various Leishmania strains. Despite this often
high degree of polymorphism, gene order is conserved between strains and species. In 1994, the Leishmania Genome Network (LGN) (http://www.ebi.ac.uk/parasites/leish.html) was set up under the auspices of the WHO to initiate a Leishmania genome sequencing project, and L. major MHOM/IL/81/Friedlin (LmjF) was selected as the reference strain. This genome is ~33.6 Mb in size, with chromosomes ranging from ~0.3 to 2.8 Mb (15). A cosmid library of 9216 clones (~9-fold genome coverage), constructed in the shuttle vector cLHYG (16), was used for a first generation cosmid contig map of the entire genome (17). The sequence of the smallest Leishmania chromosome, chromosome 1 (285 kb) revealed 79 potential protein-coding genes (1). Remarkably, the genes on this chromosome were organized into two large polycistronic units, with the first 29 genes on one DNA strand and the remaining 50 genes on the other, oriented such that their mRNAs are transcribed divergently towards the telomeres. The 257 kb informational region of chromosome 1 was flanked by telomeric and sub-telomeric sequences differing significantly in size (~29 kb) between the chromosome 1 homologs (18).

We report here the complete nucleotide sequence and analysis of the third smallest chromosome in L. major Friedlin, chromosome 3, which has a size of 384.5 kb and contains 95 putative protein-coding genes.

**MATERIALS AND METHODS**

**Cosmid mapping**

A genomic library was constructed in the shuttle cosmid cLHYG (16) using partial Sau3A1-digested genomic DNA from L. major MHOM/IL/81/Friedlin and 9216 clones were picked, arrayed and transferred to nylon membranes (17). These filters were first hybridized with a chromosome 3/2b DNA probe obtained from clamped homogeneous electric field gel electrophoresis and subsequently with a chromosome 2a DNA probe. Hybridizations were done in 1 M NaCl, 1% SDS, 100 µg/ml denatured salmon sperm DNA for 18 h at 65°C. Filters were washed twice for 20 min in 2× SSC, 1% SDS at 65°C and twice for 20 min in 0.1× SSC, 0.2% SDS at 65°C. Clones that hybridized with the chromosome 3/2b probe but not with the chromosome 2a probe were deemed to be likely to contain sequence from chromosome 3 and were initially mapped by cosmid fingerprinting (17). The precise map locations of these clones were subsequently determined by comparison of their end sequences with the consensus sequence obtained from selected cosmids that were sequenced using the shotgun method described below.

**Cosmid shotgun sequencing**

Cosmid DNA (from L952, L7535, L4625, L622, L6202, L5204, L1559, L3561, L5801, L509, L6910, L6290, L505, L7472, L712, L7234 and L3223) was sheared by sonication, repaired using the Klenow fragment of Escherichia coli DNA polymerase and T4 polynucleotide kinase and fractionated by agarose gel electrophoresis. Fragments of 0.8–2.0 kb were purified and ligated to SmaI-digested, phosphatase-treated M13mp18 RF DNA. Clones obtained from each ligation were screened for the presence of insert (white color in the presence of isopropyl β-D-thiogalactoside). DNA was prepared by PEG precipitation and templates sequenced using dye Terminator and dyePrimer chemistry with primer ~21M13 (5'-GTAAAA-CGACGGCCAGT) and ~21M14 (5'-CGATGATAAGCTGTCAAACATG). Generally 600–1000 sequencing reactions were performed during the shotgun phase before assembly using the PHRED/PHRAP/CONSED software suite (19). Gaps and regions of poor sequence quality were filled by sequencing existing clones with different chemistries and/or different primers or by sequencing PCR products amplified from the cosmid DNA. Sequences from several clones containing telomeric and sub-telomeric sequences, obtained by using a PCR-based method (20), overlapped with both ends of the chromosome 3 sequence and were included in the final chromosome 3 consensus. The entire sequence has been deposited in the GenBank sequence database with accession no. AC125735.

**Southern blot analysis**

Genomic DNA preparation and Southern blot analyses were performed as described previously (21). LmjF genomic DNA was digested with restriction enzymes, separated by 0.8% agarose gel electrophoresis and transferred to Hybond N+ membranes (Amersham) by capillary action. For use as a probe, a 560 bp fragment containing part of the ‘left-most’ gene on chromosome 3 (Chr3_0010) was PCR amplified with oligonucleotides Chr3_0010-1 (5'-ATCCGAGAGATTGCGACGTGGCAC) and Chr3_0010-2 (5'-TCTGGAGGTCGGACATTGGAGAA). A 964 bp fragment corresponding to the ‘right-most’ gene on chromosome 3 (Chr3_0890) was amplified with the oligonucleotides Chr3_0890-1 (5'-TT CCTACGCCTAGTCCG) and Chr3_0890-2 (5'-CAA TACGACGGTCCAGATGG). Fragments were labeled with [α-32P]dCTP using the High Prime labeling system (Amersham).

**Clamped homogeneous electric field analysis**

Agarose blocks containing 107 promastigotes from LmjF were prepared as previously described (22). Chromosome-sized DNA was separated from these blocks by clamped homogeneous electric field (CHEF) gel electrophoresis (Bio-Rad Mapper system) in a 1% SeaKem LE agarose gel with 1× TBE buffer. Conditions to separate DNA in the range of 150–600 kb utilized a linear switch time gradient of 15.02–44.48 s at 6 V/cm for 23 h 22 min at 14°C. The gel was transferred to a Hybond N+ membrane (Amersham) by capillary action. The Chr3_0010 fragment described above was used as a probe.

**Nuclear run-on assays**

These experiments were performed as described elsewhere (23). Briefly, nuclei were isolated from 2 × 10⁸ LmjF promastigotes by washing twice in PBS, resuspending in 10 ml ice-cold lysis buffer (10 mM Tris–HCl pH 7.5, 3 mM CaCl₂, 2 mM MgCl₂) and adding NP-40 to a final concentration of 0.5%. Cells were transferred to a Dounce homogenizer and broken with 40 strokes; the nuclei were collected by centrifugation (1400 g) and washed once with lysis buffer. To elongate nascent RNA, nuclei were resuspended in 100 µl of 100 mM Tris–HCl (pH 7.5), 25% glycerol, 0.15 mM SDS at 65°C. Filters were washed twice for 20 min in 2× SSC, 1% SDS at 65°C and twice for 20 min in 0.1× SSC, 0.2% SDS at 65°C. Clones that hybridized with the chromosome 3/2b probe but not with the chromosome 2a probe were deemed to be likely to contain sequence from chromosome 3 and were initially mapped by cosmid fingerprinting (17). The precise map locations of these clones were subsequently determined by comparison of their end sequences with the consensus sequence obtained from selected cosmids that were sequenced using the shotgun method described below.

These experiments were performed as described elsewhere (23). Briefly, nuclei were isolated from 2 × 10⁸ LmjF promastigotes by washing twice in PBS, resuspending in 10 ml ice-cold lysis buffer (10 mM Tris–HCl pH 7.5, 3 mM CaCl₂, 2 mM MgCl₂) and adding NP-40 to a final concentration of 0.5%. Cells were transferred to a Dounce homogenizer and broken with 40 strokes; the nuclei were collected by centrifugation (1400 g) and washed once with lysis buffer. To elongate nascent RNA, nuclei were resuspended in 100 µl of 100 mM Tris–HCl (pH 7.5), 25% glycerol, 0.15 mM SDS at 65°C.
spermine, 0.5 mM spermidine, 2 mM DTT, 40 U RNAsin (Promega), 2 mM MgCl₂, 4 mM MnCl₂, 50 mM NaCl, 50 mM KCl, 2 mM ATP, 2 mM GTP, 2 mM UTP, 10 μM CTP and 300 μCi of [α-32P]CTP (3000 Ci/mmol; Amersham). The incubation was carried out for 6 min at 26°C, after which DNase I (10 U) was added. Incubation was continued for 5 min at 37°C, then stopped by the addition of 100 μl of 10 mM Tris–HCl (pH 7.5), 10 mM EDTA, 1% SDS and 100 μg/ml proteinase K. After 5 min incubation at 37°C, RNA was extracted with phenol-chloroform and separated from free nucleotides by G-50 Sephadex chromatography. Labeled nascent RNA was hybridized to Hybond filters (Amersham) containing dots of 1 μg of single-stranded M13 DNA. Hybridization was performed for 48 h at 50°C in 50% formamide, 5× SSC, 0.2% SDS, 4× Denhardt’s reagent and 100 μg/ml salmon sperm DNA. Post-hybridization washes were carried out in 0.1× SSC and 0.1% SDS at 65°C. In the assays carried out in the presence of UV light, promastigotes (in a total volume of 15 ml) were irradiated in Petri dishes, with agitation, in a Stratalinker UV crosslinker (Stratagene). After irradiation, cells were incubated for at least 1.5 h at 28°C to allow the clearing of RNA polymerases engaged prior to irradiation.

Sequence analysis

The G+C content of the entire chromosome 3 sequence was calculated using the sequence and annotation viewer software ARTEMIS, version 4 (24), as was the overall G+C content of coding sequences. The REPEATMASKER program (A.F.A. Smit and P. Green, unpublished data) was used to search for simple repeats and low complexity regions within the sequence. The GCG programs WINDOW andSTATPLOT (25) were used to identify pyrimidine tracts within the chromosome 3 sequence. The results were then parsed into a suitable format for import into ARTEMIS, where their locations with respect to predicted protein-coding regions were identified.

The percentage of cumulative GC skew for the whole chromosome sequence was calculated with window length (L) = 10 kb and sliding window (∆L) = 1 kb using the expression:

\[
\%\text{GC skew} = \Sigma (\delta_G - \delta_C) \times 100/(\delta_G + \delta_C),
\]

where \(\delta_G\) and \(\delta_C\) are the counts of nucleotides G and C in the given window length.

Similarly, purine excess for the whole chromosome was calculated using the expression:

\[
P_{\text{excess}} = \Sigma (\delta_{\text{purine}} - \delta_{\text{pyrimidine}}),
\]

where \(\delta_{\text{purine}}\) and \(\delta_{\text{pyrimidine}}\) are the sum of nucleotides (A,G) and (C,T) with L and ∆L as above.

The consensus sequence for chromosome 3 was examined for putative protein-coding open reading frames (ORFs) using a combination of gene prediction techniques, combined in our Machine Automated Gene Identification (MAGI) software (25a). The methods included GLIMMER 2.0 (26) (trained with previously identified protein-coding genes from LmjF chromosomes 1 and 4), in-house adaptations of TESTCODE (27), CODONUSAGE (28) and GENESCAN (29). Putative genes predicted by at least three of these four methods, along with graphical representations of the TESTCODE and GENESCAN statistics, were exported from MAGI and imported into the ARTEMIS annotation platform (24) for visual inspection and manual editing. The sequences were searched for tRNA genes using TRNASCAN (30).

The amino acid sequence predicted from each putative gene was used as a query sequence for local BLASTP and TBLASTN (31) searches (through ARTEMIS) of the GenBank non-redundant protein and a kinetoplastid-specific nucleotide database, respectively. These databases were created from sequence collections downloaded by ftp from ftp.ncbi.nih.gov/blast/db/ and ftp.ebi.ac.uk/pub/databases/parasites/blastdbs/. Generally, hits with BLAST scores of >50 and e-values of \(<1 \times e^{-6}\) were considered potentially significant, although some exceptions were made upon visual inspection of the alignments. Each protein sequence was then searched against numerous collections of protein motifs and families (SWISS-PROT release 39.27, ProDom version 2001.2, SMART version 3.3, PROSITE release 16.46, Pfam version 6.6, PRINTS release 31.0, Domo version 2.0 and BLOCKS release 13.0). These searches were carried out using various search algorithms (http://blocks.fhcrc.org/blocks/blocks_search.html; http://hits.isb-sib.ch/cgi-bin/PFSCAN; http://dna.stanford.edu/emotif/emotif-search.html; http://www.sanger.ac.uk/Software/Pfam/). BLAST searching of the database of Clusters of Orthologous Groups of Proteins (COGs) (http://www.ncbi.nlm.nih.gov/COG/cog99nitor.html) (32) was performed simultaneously, along with identification of putative protein localization sites and transmembrane spanning regions using PSORT version II (http://psort.nibb.ac.jp/form2.html) (33) and TMHMM version 2.0 (http://www.cbs.dtu.dk/services/TMHMM-2.0/) (34). Protein domains with distinct evolutionary origins and functions were identified within the predicted genes using the NCBI Conserved Domain Database and Search Service, version 1.54 (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) (35). Gene ontology (GO) terms (36) were assigned to the predicted proteins, based on their top matches to proteins with GO annotations from SWISS-PROT and TREMBL. This analysis was carried out through the GOBlet server running at the Max Planck Institute (MPI) for Molecular Genetics (http://goblet.molgen.mpg.de/cgi-bin/OGBlet.cgi).

RESULTS

LmjF chromosome 3 was sequenced from a set of overlapping cosmids and smaller clones. A contig map of these cosmids was constructed using a combination of cosmid fingerprinting, hybridization with chromosome 3-specific probes and cosmend sequencing, as previously described (21). Seventeen cosmid clones, representing a tile-path for most of chromosome 3, were selected for shotgun sequencing (Fig. 1). In addition, 11 clones containing telomeric and sub-telomeric sequences were prepared using a PCR-based method (20), since these sequences were not present in this cosmid library. The sequences obtained from the cosmid clones and telomeric clones were assembled into a final consensus sequence of 384,518 bp.

The consensus sequence from the ‘left’ end of chromosome 3 has 1165 bp 5′ to the first protein-coding gene. The distance from the last protein-coding gene at the ‘right’ end to the end of the available sequence is 2283 bp. The consensus sequence
terminated in 40 nt of telomere hexamer repeat (THR) sequence (CCCTAA) at the ‘left’ end and 520 nt of the THR repeat at the ‘right’ end. This sequence is found at the telomeres of all Leishmania chromosomes (21,37). No other telomere-associated (TAS) or sub-telomeric (LST) repeats were found adjacent to the THR repeats in chromosome 3, unlike in other Leishmania chromosomes (18). Genomic Southern analysis using a number of different restriction enzymes and probes from each telomere detected two major fragments in each case (Fig. 2a and b). The presence of minor bands in some digests at the ‘right’ telomere is most likely due to incomplete digestions. The different fragment sizes observed presumably reflect a small size difference between chromosome homologs in the size of the telomeric regions, and indicate that a further 800 or 1710 bp (depending upon the homolog) of unsequenced DNA is located at the ‘left’ end of the chromosome 3 sequence. It is likely that this additional sequence is composed largely of THR repeats, although the presence of TAS-A and/or TAS-B sequences (18) cannot be ruled out. Similarly, a further 1180 or 2280 bp of unsequenced DNA is located at the ‘right’ end. The presence of a third, less intense, band with the Chr3_0010-1 probe, which is not due to incomplete digestion, suggests further heterogeneity in the length of the telomeric sequences at the ‘left’ end of one of the chromosome 3 homologs. Southern blot analysis of CHEF gel-separated LmjF chromosomal DNA with this probe showed hybridization only to chromosome 3 (Fig. 2c), indicating that the additional restriction fragment is not likely due to cross-hybridization with a fragment elsewhere in the genome. These results indicate that LmjF chromosome 3 has the smallest combined telomeric and sub-telomeric capping region identified in any trypanosomatid species to date.
Thus, the total size of the chromosome 3 homologs ranges from 387.1 to 389.5 kb. This is consistent with measurements from PFGE separations of LmjF chromosomes, where both chromosome 3 homologs co-migrate together with chromosome 2b, indicating that the size difference between the chromosome 3 homologs is small (15).

Analysis of the chromosome 3 sequence revealed 1154 ORFs larger than 225 nt (75 amino acids) in length. A single gene encoding a tRNA^{Lys}_{CUU} was identified using TRNASCAN. Putative protein-coding ORFs were identified using a combination of gene prediction techniques (GLIMMER 2.0, CODONUSAGE, TESTCODE and GENESCAN) as well as manual inspection in ARTEMIS. These analyses identified 95 putative protein-coding ORFs (Fig. 3). The gene density is 1 gene per 4.0 kb and the mean size of these predicted protein-coding ORFs is 2517 bp, with a range of 318 (Chr3_0870) to 9342 bp (Chr3_0690); the predicted protein-coding sequence accounts for ~60.8% of chromosome 3. No intron sequences were detected in these genes, as is typical for the Trypanosomatidae. The average G+C content for sequence predicted to be coding is 65.13%, which is somewhat higher than that of the genome in general (62.5%).

Pyrimidine-rich tracts provide part of the processing signals for the addition of the 39 nt spliced leader (or mini exon) sequence at the 5’ end of all mRNAs and for 3’ polyadenylation (41). Tracts of 10 pyrimidines or more were identified in all but two intergenic regions on chromosome 3 (data not shown). The size distribution in these regions was between 9 and 29 nt (mean length 24). Pyrimidine tracts of this size range were rare within the predicted protein-coding regions (fewer than half of the ORFs contained tracts of 10 nt or longer). The mean tract size in coding regions was 13 pyrimidines (maximum 24).

Interspersed and simple tandem repeats in the finished sequence were masked and a combination of BLASTX/BLASTP/TBLASTN database searches carried out in order to assign each of the 95 putative protein-coding genes on chromosome 3 to 13 different categories based on their putative biological function (Table 1). In parallel, a number of databases containing collections of biologically significant sites and patterns were queried with the sequence of each putative protein. Protein localization signals and probable transmembrane regions were also identified. Out of the 95 putative protein-coding genes identified on chromosome 3, 57 were predicted to be soluble, 17 were predicted to have a single transmembrane domain and 20 were predicted to have two or more transmembrane domains. Of this latter group, 12 were predicted to be localized to the plasma membrane and five to the endoplasmic reticulum.

The predicted amino acid sequences from 29 of the chromosome 3 genes (30.5%) showed sufficiently high sequence similarity to previously identified proteins to be assigned putative specific functions. The remaining 69.5% of genes remain unclassified. This percentage of unclassified genes is similar to that currently found in the whole LmjF genome project (70%). Within this category, 47% of the genes...
are predicted to encode proteins with similarity to proteins with unknown function in other organisms or show insufficient similarity (i.e. only short motifs) to allow functional assignment, 31% show similarity only to genes in other trypanosomatid species and 22% show no similarity to previously identified protein sequences.

Most of the functional protein classifications identified by the LGN are represented on chromosome 3, with the exception of proteins predicted to be involved in intracellular trafficking or cell rescue, defense, death and aging (Table 1). The numbers of genes predicted to encode for proteins with these functions within the whole genome are low in general. Of the 29 proteins whose functions have been determined, six are homologous to proteins involved in transport and a further six are homologous to proteins involved in metabolism, three of these being involved in amino acid metabolism. Proteins

Table 1. Functional classification of *L.major* chromosome 3 genes

<table>
<thead>
<tr>
<th>Category</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell growth, division and DNA synthesis</td>
<td></td>
</tr>
<tr>
<td>Putative DNA primase large subunit dp58</td>
<td>5</td>
</tr>
<tr>
<td>Helicase 1 (Chr3_0490)</td>
<td></td>
</tr>
<tr>
<td>Helicase 2 (Chr3_0610)</td>
<td></td>
</tr>
<tr>
<td>Serine-threonine protein kinase 4 (Chr3_0170)</td>
<td></td>
</tr>
<tr>
<td>RNA-binding protein (Chr3_0430)</td>
<td></td>
</tr>
<tr>
<td>Cell rescue, defense, death and aging</td>
<td></td>
</tr>
<tr>
<td>Cellular organization/biogenesis</td>
<td></td>
</tr>
<tr>
<td>Ankyrin repeat protein 1 (Chr3_0390)</td>
<td>0</td>
</tr>
<tr>
<td>Energy generation</td>
<td></td>
</tr>
<tr>
<td>Quinone oxidoreductase (Chr3_0480)</td>
<td>0</td>
</tr>
<tr>
<td>Intracellular trafficking</td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td></td>
</tr>
<tr>
<td>Long chain fatty acyl CoA synthetase (Chr3_0180)</td>
<td></td>
</tr>
<tr>
<td>2-Aminoethylphosphonate/pyruvate aminotransferase (Chr3_0030)</td>
<td></td>
</tr>
<tr>
<td>Inorganic pyrophosphatase (Chr3_0820)</td>
<td></td>
</tr>
<tr>
<td>6-1-Pyrole-5-carboxylate dehydrogenase (Chr3_0160)</td>
<td></td>
</tr>
<tr>
<td>δ-3-Phosphoglycerate dehydrogenase (Chr3_0020)</td>
<td></td>
</tr>
<tr>
<td>Multicopper oxidase 1 (Chr3_0860)</td>
<td></td>
</tr>
<tr>
<td>Protein destination</td>
<td></td>
</tr>
<tr>
<td>26S Protease regulatory ATPase subunit</td>
<td>3</td>
</tr>
<tr>
<td>Protein arginine methyltransferase (Chr3_0500)</td>
<td></td>
</tr>
<tr>
<td>Cytochrome c oxidase assembly protein (Chr3_0080)</td>
<td></td>
</tr>
<tr>
<td>Protein synthesis</td>
<td></td>
</tr>
<tr>
<td>Elongation initiation factor 2, α subunit (Chr3_0890)</td>
<td></td>
</tr>
<tr>
<td>Ribosomal protein L38 (Chr3_0200)</td>
<td></td>
</tr>
<tr>
<td>60S Acidic ribosomal protein P1_2 (Chr3_0350)</td>
<td></td>
</tr>
<tr>
<td>60S Acidic ribosomal protein P1_1 (Chr3_0345)</td>
<td></td>
</tr>
<tr>
<td>Signal transduction</td>
<td></td>
</tr>
<tr>
<td>Serine-threonine protein kinase 2 (Chr3_0280)</td>
<td></td>
</tr>
<tr>
<td>Serine-threonine protein kinase 3 (Chr3_0090)</td>
<td></td>
</tr>
<tr>
<td>Structural RNA</td>
<td></td>
</tr>
<tr>
<td>tRNA^{CAU} (Chr3_tRNA_Lys01)</td>
<td>1</td>
</tr>
<tr>
<td>Transcription and RNA processing</td>
<td></td>
</tr>
<tr>
<td>U2AF23/25 spliceosome component (Chr3_0155)</td>
<td></td>
</tr>
<tr>
<td>Transport</td>
<td></td>
</tr>
<tr>
<td>Lysine-arginine-ornithine transport system kinase 1 (Chr3_0510)</td>
<td></td>
</tr>
<tr>
<td><em>Leishmania</em> unknown permease 1 (Chr3_0330)</td>
<td></td>
</tr>
<tr>
<td>Phosphate-repressible phosphate permease 1 (Chr3_0410)</td>
<td></td>
</tr>
<tr>
<td>ABC transporter protein 1 (Chr3_0130)</td>
<td></td>
</tr>
<tr>
<td>Choline transporter protein 1 (Chr3_0060)</td>
<td></td>
</tr>
<tr>
<td>Tubulin-specific chaperone (Chr3_0680)</td>
<td></td>
</tr>
<tr>
<td>Unclassified</td>
<td></td>
</tr>
<tr>
<td>Chr3_0115; Chr3_0140; Chr3_0150; Chr3_0205; Chr3_0220; Chr3_0290; Chr3_0360; Chr3_0370; Chr3_0470; Chr3_0640; Chr3_0650; Chr3_0720; Chr3_0770; Chr3_0850.</td>
<td>65</td>
</tr>
<tr>
<td>No homology/motifs</td>
<td></td>
</tr>
<tr>
<td>Chr3_0010; Chr3_0015; Chr3_0040; Chr3_0100; Chr3_0110; Chr3_0210; Chr3_0240; Chr3_0240; Chr3_0270; Chr3_0300; Chr3_0310; Chr3_0312; Chr3_0315; Chr3_0340; Chr3_0400; Chr3_0420; Chr3_0460; Chr3_0510; Chr3_0590; Chr3_0600; Chr3_0630; Chr3_0670; Chr3_0700; Chr3_0710; Chr3_0730; Chr3_0760; Chr3_0780; Chr3_0790; Chr3_0800; Chr3_0830; Chr3_0840; Chr3_0880.</td>
<td>32</td>
</tr>
<tr>
<td>Uninformative blast homology/motifs</td>
<td></td>
</tr>
<tr>
<td>Chr3_0050; Chr3_0090; Chr3_0120; Chr3_0175; Chr3_0180; Chr3_0250; Chr3_0260; Chr3_0380; Chr3_0440; Chr3_0485; Chr3_0520; Chr3_0540; Chr3_0550; Chr3_0560; Chr3_0580; Chr3_0620; Chr3_0660; Chr3_0740; Chr3_0750; Chr3_0870.</td>
<td>20</td>
</tr>
<tr>
<td>Shared with other Kinetoplastidae</td>
<td></td>
</tr>
<tr>
<td>Total for chromosome 3</td>
<td>96*</td>
</tr>
</tbody>
</table>

*Including the structural RNA Chr3_tRNA_Lys01
involved in cell growth, division and DNA synthesis and in protein synthesis are also relatively common on this chromosome (accounting in total for another nine genes). Two DEAD/DEAH box helicases have been identified, as have two unrelated permeases and three ribosomal proteins.

Three leucine-rich repeat (LRR) proteins, which show no significant identity to one another at the protein level, are encoded on chromosome 3 (Chr3_0010, Chr3_0100 and Chr3_0780). The functions of these proteins are unknown, but proteins containing these characteristic LRR motifs have been identified in a wide variety of species and are associated with widespread differing functions (42). It has been suggested that LRRs may mediate protein–protein interactions, as well as cellular adhesion (43). Other identified functions of LRR-containing proteins include binding to enzymes (44) and vascular repair (45). Two glycoproteins of Leishmania, the parasite surface antigen-2/gp46 (PSA-2) and proteophosphoglycan (PPG1) are known to contain LRRs (46). These findings have led to the recent suggestion that proteins containing these motifs may be involved in interactions with the mammalian host cell and/or insect vector (47). One LRR-containing protein from chromosome 3 (Chr3_0010) also contains a RING finger motif and shows greatest similarity to ESAG8/T-LR, which is encoded in several VSG gene expression sites of Trypanosoma brucei (48). While the function of ESAG8 is unknown, it has recently been shown to interact with a protein with a possible role in regulation of mRNA stability (49).

A number of gene duplication events are apparent on chromosome 3. It is estimated that between 5 and 10% of the genes in L. major Friedlin occur in more than one copy, either as gene duplications or as members of multigene families (50). The neighboring genes, Chr3_0300, Chr3_0310, Chr3_0312 and Chr3_0315, show varying degrees of similarity at both the DNA and amino acid levels. At the amino acid level, Chr3_0315 and Chr3_0310 share 67% identity, whilst Chr3_0312 and Chr3_0310 share 39%; the remaining pairings have identities of ~20%, with the exception of Chr3_0315 and Chr3_0300, which show only 9% identity. Thus, it appears that two separate duplication events occurred to give rise to these four genes. Interestingly, whilst all of these proteins are predicted to belong to a group of uncharacterized, hydrophobic, evolutionarily related membrane proteins (GPR1/FUN34/YAAH) from fungi, bacteria and archaea (51,52), only Chr3_0300, Chr3_0310 and Chr3_0315 were predicted to contain multiple transmembrane domains. Ribosomal protein P1 copies 1 and 2 (Chr3_0345 and Chr3_0350) are identical at the amino acid level (only a single synonymous nucleotide substitution separates them at the DNA level), suggesting that the duplication event that gave rise to them may have occurred relatively recently. The level of identity at both the protein (26%) and nucleotide (13%) level is lower for Chr3_0560 and Chr3_0580, perhaps suggesting that this duplication is more ancient and that these two proteins may now have divergent functions.

The 95 putative protein-coding genes are organized into two large polycistronic clusters of 65 and 29 genes, plus a single gene located at one telomere (Fig. 3). The two clusters are organized convergently, i.e. transcription of the mRNAs is orientated away from each telomere. A tRNA gene is located within the convergent strand switch region at the boundary between the clusters. A single gene at the ‘left’ telomere is transcribed towards the telomere, i.e. divergently from the neighboring gene cluster. Nuclear run-on studies of chromosome 3 (Fig. 3a) indicate that transcription of chromosome 3 mirrors the gene organization pattern, since the hybridization signal for the gene-coding strand is ~10-fold higher than that for the non-coding strand. Thus, transcription of chromosome 3 appears to be polycistronic and initiates upstream of each gene cluster, as seen for LmjF chromosome 1 (23).

REPEATMASKER, a search program for simple repeats and low complexity regions, identified 257 mono- to hexanucleotide microsatellite sequences, including the telomeric hexamer repeats, comprising ~3% of chromosome 3. Of the microsatellite sequences identified, 44% were GC-rich, 4% were AT-rich and 52% contained equal numbers of C/G and A/T nucleotides. CA/GG dinucleotide repeats were the most abundant class of microsatellites, accounting for 1% of the chromosome. A further 0.5% of chromosome 3 was GC-rich low complexity sequence and 2% was made up of other classes of tandemly repetitive DNA, including minisatellite sequences (repeats with motifs of 6–18 nt). No putative transposable elements were identified in this LmjF sequence.

GC skew is a statistical method to measure the strand-specific over-representation of guanine and the inflection point on a plot of GC skew versus nucleotide position corresponds to the position of the putative origin or terminus of DNA replication in a number of organisms (53). A non-random distribution of nucleotide bias was found for chromosome 3 using analyses similar to that previously described (54). The maximum value of cumulative GC skew identified coincides with the region between the single gene at the ‘left’ end of the chromosome and the polycistronic cluster of neighboring genes (Fig. 3b). Purine excess analysis has also been used to identify the putative origin or terminus of DNA replication in a number of bacterial species (55). The maximum value for purine excess is also located at this strand-switch point. The minimum GC skew and purine excess values coincide with the strand-switch region between the two polycistronic gene clusters, where the tRNA is located. A second maximum for GC skew purine excess is located upstream of the THR sequences at the ‘right’ end of chromosome 3.

**DISCUSSION**

This paper reports the complete sequence of LmjF chromosome 3. This sequence adds to the published large-scale contiguous finished sequence from protozoan parasite genomes, which currently consists of a draft sequence of the entire Plasmodium falciparum genome (56), one chromosome from L. major (21) and a region of chromosome 3 from Trypanosoma cruzi (57). Ninety-five protein-coding genes were identified on this chromosome, including proteins belonging to families as yet unidentified in this species. The gene density of LmjF chromosome 3 is 1 gene every 4.0 kb, which is slightly lower than the 1 every 3.3 kb observed for the informational region of LmjF chromosome 1 (21), but consistent with that observed for those LmjF chromosomes with large sections of completed sequence (chromosomes 1–6, 12–15, 19, 21–25, 27, 29, 31 and 34–36) (data not shown). In contrast, the gene density for a 75 kb portion of chromosome 21 is 1 gene every 2.4 kb (58), but the genes within this region...
are significantly smaller on average than those on chromosome 1, and probably reflect a cluster of smaller genes. Clusters of larger, more dispersed genes are also seen elsewhere in the LmjF genome. The gene density is also comparable to that obtained from *T. cruzi* (1 per 4.5 kb) (57) and *P. falciparum* (1 per 4 kb) (59), but is somewhat higher than that of *Caenorhabditis elegans* (1 gene per 5 kb) (60) and lower than that of *Saccharomyces cerevisiae* (1 gene per 2 kb) (61).

As first reported for chromosome 1 (21), the genes on LmjF chromosome 3 are organized into large polycistronic clusters. Similar organization of genes in *T. cruzi* and *T. brucei* (57) indicates that this pattern is common in trypanosomatids. However, in contrast to chromosome 1, chromosome 3 contains two large convergent polycistronic gene clusters and a single divergently transcribed gene. Whilst genes with related functions are clustered in polycistronic transcription units in many prokaryotic species, thus facilitating cooperation (62), no such organization is the case in LmjF chromosome 3 or chromosome 1. The polycistronic units of LmjF are much larger than those found in these bacterial operons and contain genes with apparently unrelated functions. This suggests that the basis for this organization into large polycistronic clusters is unrelated to gene function.

LmjF chromosomes 1 (54) and 3 have GC skew and purine excess patterns that mirror the polycistronic gene clusters. The effects of selective pressure on events associated with replication, transcription and DNA repair are thought to influence base composition (63,64). Replication-associated mutations tend to skew gene locations to the lagging DNA strand (65). In addition, errors associated with transcription may also skew the base composition of the different DNA strands (55). However, the nucleotide bias in LmjF is the opposite of that seen in bacteria, where there is a positive correlation between gene orientation and GC skew, AT skew or purine excess (66). Hence, the organization of genes in clusters and the patterns of base skew and base composition seen in *Leishmania* may reflect replication and/or transcription processes. This implies that regions upstream of the clusters may contain replication origins and/or promoters.

Little is known about transcription of protein-coding genes by RNA polymerases II and III in LmjF and in other trypanosomatid species. Most protein-coding genes in these organisms are transcribed polycistronically (4) and regulation of their gene expression is believed to be primarily post-transcriptional, occurring at the levels of trans-splicing, polyadenylation, mRNA stability, translation or protein stability (4), suggesting that trypanosomatids may have relatively few polymerase II promoters. Transcriptional analysis of LmjF chromosome 1 using strand-specific nuclear run-on assays shows that transcription preferentially initiates within a <100 bp region in the strand-switch region between the two gene clusters, from which transcription appears to proceed bidirectionally towards each telomere (23). Nuclear run-on analysis of chromosome 3 indicates that transcription is similarly consistent with the pattern of gene organization (Fig. 3a). The results obtained using UV-irradiated nuclei suggest that transcription initiates in three different regions: between the divergently organized left-most gene and large left gene cluster, telomeric to the right-most gene cluster and upstream of the tRNA gene that separates the two gene clusters (Martinez-Calvillo *et al.*, submitted for publication). RNA polymerase II appears to mediate transcription at the first two sites, while the last is carried out by polymerase III. Transcription of both strands appears to terminate within the tRNA gene region (Martinez-Calvillo *et al.*, submitted for publication).

The sequence of the region between the two polycistronic clusters on chromosome 3 predicts significant curvature and stem-loop structures (58). Curvature has been associated with transcription promoter activity (67) as well as replication (68) and centromere functions (69). No centromeric or replication origin consensus sequence of any kinetoplastid has been identified to date. In a recent study, targeted replacement of the strand-switch region on two copies of chromosome 1 did not prevent normal replication and segregation of these chromosomes (70). This may be taken to indicate that it does not constitute an origin of replication or centromere, but it is possible that trans-complementation from the strand-switch region on the remaining chromosome 1 copy could be maintaining the other chromosomes lacking origins.

The structure of telomeric and sub-telomeric regions in a wide variety of organisms is highly plastic, often varying significantly in both size and composition between chromosomes within a single genome (71). Large differences are seen in the size of the sub-telomeric sequence between LmjF chromosome 1 homologs (18), due to the differences in the number of copies of simple sequence repeats. It had previously been shown that at least one telomere of chromosome 3 lacked the sub-telomeric repeats found in other LmjF chromosomes (72). The complete sequencing of chromosome 3 shows that there are only modest (1–2 kb) differences in the size of the telomeric sequences between homologs and that both ends lack sub-telomeric repeats. The role of sub-telomeric repeats in *Leishmania* is an open question. They are likely to play a role in the maintenance of genome stability by protecting the non-repetitive chromosomal termini from degradation, end-to-end fusion and rearrangement (73,74). Rearrangements in telomeric repeat sequences have been linked to immune evasion through alteration of immunogenic epitopes on the cell surface in other organisms (75,76), though this has not yet been observed in *Leishmania*. It has recently been suggested that the 272 bp repeats in the sub-telomeric region of chromosome 1a have centromeric function, based upon analysis of a version of chromosome 1a in which all sequence other than the ‘right-hand’ sub-telomeric and telomeric repeats was deleted through homologous recombination (77). Similarly, analysis of chromosome fragmentation experiments on amplified small linear chromosome in *Leishmania donovani* suggested that the region responsible for mitotic stability contained homopolymer tracts (78). The absence of sub-telomeric repeats on LmjF chromosome 3 indicates that they are not a requirement for chromosome stability, re-opening the question of their biological function in other chromosomes.

Only 31% of the chromosome 3 genes could be assigned function on the basis of homology to proteins from other organisms. These proteins span a wide range of functional classes, as detailed in Table 1. Approximately 33% show homology to unclassified proteins in other species. Substantial numbers (36%) of chromosome 3 genes show no sequence similarity to proteins from organisms other than related
trypanosomatid species. These genes may be specific to the trypanosomatids or may have diverged sufficiently as to have no significant sequence similarity to their functional homologs in other species. Since the total number of genes in LmjF is likely to be about 8600 (79), it is reasonable to infer that completion of the entire Leishmania genome sequence may identify approximately 3000 genes with potentially parasite-specific functions.

Completion of the sequence of chromosome 3 increases knowledge of the molecular biology of these parasites, contributing to the overall understanding of genome structure and comparative biology/evolution of this and related species. It has led to the identification of 95 new genes with key cellular functions that may be potential drug targets, vaccine candidates or useful for diagnostics or for vaccine development. In addition, study of LmjF chromosomes is valuable for future analysis of many aspects of basic molecular biology.

ACKNOWLEDGEMENTS

This work was supported by a PHS grant (R01 AI40599) from the National Institutes of Health to K.D.S. and post-doctoral fellows from the International Training and Research in Emerging Infectious Diseases (ITREID) program to S.M.-C. and G.A.

REFERENCES


