ELSEVIER

Molecular & Biochemical Parasitology 113 (2001) 337–340

MOLECULAR & BIOCHEMICAL PARASITOLOGY

www.parasitology-online.com.

Short communication

A survey of the *Leishmania major* Friedlin strain V1 genome by shotgun sequencing: a resource for DNA microarrays and expression profiling☆

Natalia S. Akopyants [a], Sandra W. Clifton [b], John Martin [b], Deana Pape [b], Todd Wylie [b], Li Li [c], Jessica C. Kissinger [c], David S. Roos [c], Stephen M. Beverley [a,*]

[a] *Department of Molecular Microbiology, Washington University School of Medicine, St. Louis MO, USA*
[b] *Genome Sequencing Center, Washington University, St. Louis MO, USA*
[c] *Department of Biology, University of Pennsylvania, Philadelphia PA, USA.*

Received 4 October 2000; accepted 19 January 2001

Keywords: Genome survey sequence (GSS); Expression profiling; Gene discovery; Transcriptional regulation; Chromatin

Leishmania are pathogenic trypanosomatid protozoans responsible for a diverse spectrum of human diseases. These parasites exhibit a dimorphic life cycle, consisting of extracellular promastigotes that reside within the midgut of the sand fly vector, and intracellular amastigotes that reside within the phagolysosome of host macrophages. How *Leishmania* differentiate, survive and multiply in such inhospitable environments is a challenging question relevant to disease control.

In the last 10 years, genetic tools have been developed which permit a variety of modifications to the parasite genome, using expression vectors and gene knockout technology [1]. A genome project has been initiated, and the sequence of several chromosomes has recently been completed [Refs. [2] [3] [4]; and http://www.ebi.ac.uk/parasites/leish.html]. The combination of extensive genome sequence information and reverse genetic tools permits powerful new approaches for understanding gene function in this parasite. Experimenters must now confront the question of how best to choose amongst the ~10 000 proteins encoded by the parasite genome for further study. This challenge is compounded by the fact that most *Leishmania* genes show no obvious sequence similarity to those encoding any known proteins [Ref. 4, this work].

One criterion for prioritizing genes for further study is based on the concept of expression profiling, in which the level of mRNA abundance is determined for thousands of genes simultaneously [5]. This may be accomplished by hybridization of total labeled mRNA (or cDNA) preparations to arrays of DNAs on glass slides, followed by quantitation of specific hybridization by fluorescence [6,7]. Comparative analysis of hybridizations using different preparations permits, for example, the identification of genes that are expressed in a stage-specific manner. Although many routes of gene regulation do not involve changes in mRNA abundance (a key point in *Leishmania* and trypanosomatid parasites, given their reliance upon polycistronic transcriptional mechanisms [8,9]) a number of mRNAs implicated in *Leishmania* virulence have been found to exhibit differential regulation [8,10–12].

DNA microarrays typically exploit well-characterized collections of cDNAs, or completed genome sequences from which oligonucleotide probes can be synthesized directly on a chip [5]. As neither of these resources is yet available for *Leishmania*, we elected to pursue an alternative approach based upon random shotgun DNA sequences. This approach is feasible in trypanosomatid parasites due to their high gene density

☆ *Note*: The sequences reported in this work have been deposited in the GSS section of GenBank.

* Corresponding author. Tel.: +1-314-7472630; fax: +1-314-7472634.

E-mail address: beverley@borcim.wustl.edu (S.M. Beverley).

0166-6851/01/$ - see front matter
PII: S0166-6851(01)00227-4

(intergenic regions are typically <400 nt) and low frequency of introns [9,13]. Similar approaches have been successfully used to survey other parasite genomes, including *Trypanosoma brucei rhodesiense* and *Plasmodium falciparum* [14–16].

To estimate the ability of random shotgun sequence collections to identify *Leishmania major* transcripts, we considered the probability that a library of 10 000 random 1 kb sequences would identify any given transcript. Note that this question differs significantly from the problem of identifying all nucleotides, as assessed by the Lander–Waterman equation commonly used to estimate the coverage expected for recombinant clone collections [17]. Since a collection of 10 000 1 kb clones would constitute about 30% of the 34 Mb *L. major* genome, it could not possibly cover a higher percentage of the genome nucleotides. However, in Northern blot or microarray-based expression profiling, it is only necessary that the random shotgun sequences overlap enough of the transcript to serve as a useful probe.

For this calculation, we took the average gene length as 3.3 kb and the average inter-transcript spacing of 400 nt [4,9], yielding an average transcript length of 2.9 kb. Assuming uniform transcript length and spacing, no intervening sequences, and uniform gene distribution throughout the genome, the probability that a single random 1 kb insert will overlap a given transcript by at least 100 nt (permitting hybridization) can be calculated as: [2.9 kb (transcript length) + 1 kb (insert length) − 200 nt ($2\times$ required overlap)]/34 Mb (genome size) = 0.000109. The probability that at least one of N clones will recognize any given transcript is then $1-(1-0.000109)^{N}$; for 10 000 random 1 kb sequences, this yields a probability of $\sim$66%. Experimental observations support this estimate, as described below. Because such a library would provide a reasonable degree of transcript coverage for *Leishmania*, we set out to create it in a form that would be suitable for both random shotgun sequencing as well as for microarrays and expression profiling.

Genomic DNA from *L. major* strain Friedlin clone V1 (MHOM/JL/80/Friedlin) was isolated as described [18]. To minimize bias, DNA was sheared by neubilization [19]. DNAs of 1–1.5 kb were purified by agarose gel electrophoresis, their ends repaired using the Klenow fragment of DNA polymerase, and inserted into the *Eco*RV site of pZERO-2 (Invitrogen) by blunt-end ligation. DNAs were transformed into *E. coli* strain Top 10 (Invitrogen) and plated on media containing 25 μg ml^{-1} kanamycin. The pZERO-2 vector uses a positive selection scheme for inserts, and PCR amplification using flanking primers indicated that 91% (86/96) of the colonies tested contained inserts. A library of $\sim$100 000 independent clones was obtained, and 10 464 of the primary colonies were transferred to 96 well microtiter plates. This library (designated the '*L. major* GSS collection') has subsequently been re-arrayed into 384 well plates and can be obtained from Incyte Genomics Inc. (http://reagents.incyte.com/index.html ; Tel.: +1-800-4300030).

For sequencing, double-stranded plasmid DNA templates were prepared using dye terminator chemistry and AmpliTaq DNA polymerase, and single-pass sequencing was carried out using Applied BioSystems Automated 377 sequencers. After editing to remove vector contamination and low quality sequence, useful information was obtained from 65% of the GSS DNAs, including both ends of 3779 clones, and a single end from 2756 clones. A total of 10 314 sequences were submitted to the GenBank (NCBI) genome survey sequence (GSS) database (http://www.ncbi.nlm.nih.gov). This increased the amount of *Leishmania* sequence available to investigators for gene discovery purposes, adding 4.2 Mb to the 4.1 Mb of finished sequenced reported by the genome project (as of 30 September 2000).

Although GSS sequences could potentially be contaminated by kinetoplast DNAs, screening for the 13 nt universal minicircle replication origin sequence [20] identified only eight minicircle sequences (<0.1%). This probably arises from the fact that minicircles are typically less than 700 bp in *L. major* [20], while DNAs of 1–1.5 kb were selected for cloning. Twenty two GSS sequences (derived from 16 different recombinants) showed similarity to sequences present in the kinetoplast DNA maxicircle of *Leishmania tarentolae* (GenBank M10126). It is likely that some maxicircle sequences could be missed due to the occurrence of RNA editing, or sequence divergence in the variable region [20,21]. Analysis of the G/C base composition showed that most maxicircle sequences were relatively A/T-rich (<35% vs. $\sim$63% for the average GSS sequence), and several anonymous A/T-rich sequences were identified which potentially could arise from maxicircle DNA. Overall, kinetoplast DNA representation in our library was minimal (<0.3%).

Database comparisons showed that 971 of the GSS sequences (9.4%) contain various repetitive elements such as telomeric and subtelomeric repeats [4]. 7543 sequences (73.1%) lacked obvious similarity to any protein sequence in GenBank, while 1800 (17.5%) showed significant similarities. These results are summarized at the websites of the *Leishmania* Genome project (http://www.ebi.ac.uk/parasites/leish.html) and the Parasite Clustered Sequence database for *Leishmania* (http://paradb.cis.upenn.edu/leish/index.html; see below). As expected, a large number of genes showing relationship to proteins of known functions were detected, and the websites above provide various views and classifications of these. Many of these proteins are potentially relevant to studies on *Leishmania* virulence and metabolism.

In addition to the 10 314 GSS sequences reported here, 1281 genomic cosmid or PAC end sequences and 2191 ESTs have previously been deposited in GenBank, and an additional 2972 unpublished cosmid end sequences were provided by P. Myler (Seattle Biomedical Research Institute). Genomic sequence (including both GSS, cosmid and PAC end sequences) was assembled together using the cap2 software for clustering [22] yielding a total of 8894 assemblies, of which 7177 were singletons. EST sequences were also clustered together to minimize redundancy in the database [22] yielding 1,156 clusters (806 singletons). To facilitate gene identification, these assembled sequences were combined with all *Leishmania* data available from GenBank (including high throughput data from the genome project), to create a single BLAST-queryable database which can be found at http://paradb.cis.upenn.edu/leish/index.html. This website provides several analyses, including:

- BLASTN comparisons of the database against itself —identifying overlap between GSS/EST/HTG sequence data, gene duplications, and gene family members. Comparisons of cDNAs and genomic sequences may potentially reveal examples of *cis*-splicing.
- Searches for predicted Prosite motifs Ref 23, and BLASTN and BLASTX comparisons against the non-redundant GenBank database — providing putative gene identification. The results of these analyses were then combined into a text-queryable database, permitting searches for any sequences exhibiting similarity to genes of known function (i.e. proteases, nucleases, transcription factors, etc.) identified in the GenBank descriptions.

Curiously, while most typical classes of eukaryotic proteins were well represented in BLAST searches (websites noted above), no obvious homologs of the RNA polymerase II complex other than several basal core subunits were detected. We also noticed a lack of genes showing convincing relationships to eukaryotic transcription factors/activators. In contrast, our searches identified many genes predicted to encode proteins involved in maintaining chromatin structure (histones H2A, H2B and H4; GSS accession numbers: AQ847930, AQ845483 and AQ853010, respectively), or regulating the accessibility of RNA polymerase to chromatin (histone acetyl transferase and histone deacetylase; AQ846498, AQ850474, AQ847037). The same phenomenon has also been noted in scans of genome sequence data from trypanosomes (A. Günzl, personal communication). In contrast, the genome of *Saccharomyces cerevisiae* encodes hundreds of predicted transcription and initiation factors, and a comparable fraction of other eukaryotic genomes is devoted to these functions. While complete genome sequence is not yet available for any trypanosomatid, the collective representation is certainly high enough to make the lack of obvious transcription factors remarkable. To date, no RNA polymerase II promoter has yet been identified for the protein coding genes of *Leishmania* or trypanosomes [1]. Thus, the information emerging from kinetoplastid genomes suggests that genes associated with the regulation of transcriptional initiation for protein coding genes by RNA polymerase II may well be absent, or so divergent as to be presently unrecognizable by current homology search algorithms. Potentially, the reliance upon polycistronic transcription and *trans*-splicing in trypanosomatids minimizes the need for such factors.

The availability of complete sequence information for *L. major* chromosomes 1 and 3 [Ref. [4] and P. Myler, unpublished] permitted calculation of the fraction of ORFs from these chromosomes that are represented in the *L. major* GSS clone collection. Of the 79 ORFs identified on chromosome 1, 49 (62%) were identified one or more times in the *L. major* GSS clone collection. Similarly, 59 of the 93 ORFs on chromosome 3 were identified (63%), and comparable results were obtained with several other randomly chosen regions of the *Leishmania* genome (data not shown). Thus, nearly 2/3 of *Leishmania* ORFs are represented by one or more DNAs present in the GSS collection. This is in excellent agreement with the theoretical calculations for transcript coverage presented earlier (66%).

While this coverage is remarkably high, it may underestimate the ability of the GSS clone collection to identify *Leishmania* transcripts. First, only 4.2 Mb of the approx. 10 Mb of sequence arising from the GSS clone collection was determined; in a microarray hybridization experiment, the entire 10 Mb sequence would be available. Additionally, open reading frame comparisons exclude the 5′ and 3′ untranslated regions, which typically constitute $\sim 50\%$ of *Leishmania* mRNAs and likewise would be detected in microarray analysis. While it is difficult to accurately predict the true coverage of the *Leishmania* transcriptome by the GSS set on purely theoretical grounds, the factors above suggest that our estimates of 2/3 transcripts identified are if anything underestimates.

In summary, random shotgun collections provide an efficient means for generating DNAs capable of identifying a large fraction of *Leishmania* mRNAs, and are therefore well suited for expression profiling studies. While not every transcript will be identified, this approach offers several advantages over traditional arrays based on complete genome sequence information: it is relatively inexpensive to prepare DNA collections, and expression-profiling analysis need not wait for the availability of complete genomic DNA sequence. Genomic shotgun libraries may not be suitable for genomes with a low gene density, with numerous (or large) introns, or where differential splicing is extensive.

Fortunately, many protozoan parasites and pathogenic fungi resemble *Leishmania* and other trypanosomatids in their suitability for genomic shotgun-based approaches, with a high gene density coupled with small gene size (arising from a paucity of introns). Lastly, end sequencing of random DNA clones also provides an efficient means for gene discovery, avoiding the problems associated with differential abundance of mRNAs associated with EST approaches.

Analysis of the gene expression profiles during parasite development, or in response to various experimental treatments (pH, temperature, inhibitors) or genetic alterations, should allow a better understanding of *Leishmania* biology and host relationships. In turn, this will facilitate the discovery of novel gene products that could represent targets for the development of new drugs and vaccines. We are currently applying the *L. major* GSS collection described here in expression profiling experiments for these purposes.

Acknowledgements

We thank Heidi Brennecke and Elena N. Bukanova for technical assistance; Warren Ewens for assistance with statistical calculations; Christine Clayton, Deborah Dobson, George Cross and Arthur Günzl for discussions; Peter Myler for providing unpublished sequences and helpful discussions; Brian Brunk for permission to use bioinformatic clustering software tools developed as part of the *Toxoplasma* EST project; and the *Leishmania* Genome Project for posting and updating various analysis of the GSS sequences presented here. This work was supported by NIH grants AI-29646 (SMB) and AI-28724 (DSR). DSR is a Burroughs Wellcome Scholar in Molecular Parasitology.

References

[1] Clayton CE. Genetic manipulation of kinetoplastida. Parasitol Today 1999;15:372–8.
[2] Ivens AC, Smith DF. Parasite genome analysis. A global map of the Leishmania major genome: prelude to genomic sequencing. Trans R Soc Trop Med Hyg 1997;91:111–5.
[3] Ivens AC, Lewis SM, Bagherzadeh A, Zhang L, Chan HM, Smith DF. A physical map of the Leishmania major Friedlin genome. Genome Res 1998;8:135–45.
[4] Myler PJ, Audleman L, deVos T, Hixson G, Kiser P, Lemley C, Magness C, Rickel E, Sisk E, Sunkin S, Swartzell S, Westlake T, Bastien P, Fu G, Ivens A, Stuart K. Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes. Proc Natl Acad Sci USA 1999;96:2902–6.
[5] Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet 1999;21(suppl.):33–7.
[6] Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW. Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotechnol 1998;16:301–6.
[7] Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G. Making and reading microarrays. Nat Genet 1999;21:15–9.
[8] Flinn HM, Smith DF. Genomic organisation and expression of a differentially-regulated gene family from Leishmania major. Nucleic Acids Res 1992;20:755–62.
[9] LeBowitz JH, Smith HQ, Rusche L, Beverley SM. Coupling of poly(A) site selection and trans-splicing in Leishmania. Genes Dev 1993;7:996–1007.
[10] Brodin TN, Heath S, Sacks DL. Genes selectively expressed in the infectious (metacyclic) stage of Leishmania major promastigotes encode a potential basic-zipper structural motif. Mol Biochem Parasitol 1992;52:241–50.
[11] Nourbakhsh F, Uliana SR, Smith DF. Characterisation and expression of a stage-regulated gene of Leishmania major. Mol Biochem Parasitol 1996;76:201–13.
[12] Coulson RM, Connor V, Chen JC, Ajioka JW. Differential expression of Leishmania major beta-tubulin genes during the acquisition of promastigote infectivity. Mol Biochem Parasitol 1996;82:227–36.
[13] 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. A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. RNA 2000;6:163–9.
[14] El-Sayed NM, Donelson JE. A survey of the Trypanosoma brucei rhodesiense genome using shotgun sequencing. Mol Biochem Parasitol 1997;84:167–78.
[15] El-Sayed NM, Hegde P, Quackenbush J, Melville SE, Donelson JE. The African trypanosome genome. Int J Parasitol 2000;30:329–45.
[16] Hayward RE, Derisi JL, Alfadhli S, Kaslow DC, Brown PO, Rathod PK. Shotgun DNA microarrays and stage-specific gene expression in Plasmodium falciparum malaria. Mol Microbiol 2000;35:6–14.
[17] Beverley SM, Coderre JA, Santi DV, Schimke RT. Unstable DNA amplifications in methotrexate-resistant Leishmania consist of extrachromosomal circles which relocalize during stabilization. Cell 1984;38:431–9.
[18] Bodenteich A, Chissoe S, Wang YF, Roe BA. In: Techniques. Venter JC, editor. Shotgun cloning as the strategy of choice to generate templates for high-throughput dideoxynucleotide sequencing in automated DNA sequencing and analysis. London: Academic Press, 1993:42–50.
[19] Simpson L. The mitochondrial genome of kinetoplastid protozoa: genomic organization, transcription, replication and evolution. Annu Rev Microbiol 1987;41:363–82.
[20] Stuart K. RNA editing in trypanosomatid mitochondria. Annu Rev Microbiol 1991;45:327–44.
[21] Huang X. An improved sequence assembly program. Genomics 1996;33:21–31.
[22] Ajioka JW, Boothroyd JC, Brunk BP, Hehl A, Hillier L, Manger ID, Marra M, Overton GC, Roos DS, Wan KL, Waterston R, Sibley LD. Gene discovery by EST sequencing in Toxoplasma gondii reveals sequences restricted to the Apicomplexa. Genome Res 1998;8:18–28.
[23] Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. Nucleic Acids Res 1999;27:215–9.