

FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, *Plasmodium falciparum*

Junichi Watanabe*, Masahide Sasaki¹, Yutaka Suzuki¹ and Sumio Sugano¹

Department of Parasitology and ¹Department of Virology, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan

Received August 30, 2000; Revised and Accepted September 29, 2000

DDBJ/EMBL/GenBank accession nos AU086071–AU088560

ABSTRACT

FULL-malaria is a database for a full-length-enriched cDNA library from the human malaria parasite *Plasmodium falciparum* (<http://133.11.149.55/>). Because of its medical importance, this organism is the first target for genome sequencing of a eukaryotic pathogen; the sequences of two of its 14 chromosomes have already been determined. However, for the full exploitation of this rapidly accumulating information, correct identification of the genes and study of their expression are essential. Using the oligo-capping method, we have produced a full-length-enriched cDNA library from erythrocytic stage parasites and performed one-pass reading. The database consists of nucleotide sequences of 2490 random clones that include 390 (16%) known malaria genes according to BLASTN analysis of the nr-nt database in GenBank; these represent 98 genes, and the clones for 48 of these genes contain the complete protein-coding sequence (49%). On the other hand, comparisons with the complete chromosome 2 sequence revealed that 35 of 210 predicted genes are expressed, and in addition led to detection of three new gene candidates that were not previously known. In total, 19 of these 38 clones (50%) were full-length. From these observations, it is expected that the database contains ~1000 genes, including 500 full-length clones. It should be an invaluable resource for the development of vaccines and novel drugs.

INTRODUCTION

Malaria is the most devastating parasitic disease in the world; it kills two million people every year. Thus, the causative agent, *Plasmodium falciparum*, has been the first target for genome sequencing of a eukaryote pathogen. Its 30 Mb genome with an unusually high AT content (80%) is divided into 14 chromosomes, for two of which the complete nucleotide

sequence has already been deciphered (1,2) and the rest will be determined within a few years.

This parasite has a quite complex life cycle: a sexual stage in mosquitoes and an asexual stage in humans, with proliferation first in hepatocytes and then in erythrocytes. Analysis of gene expression at different stages is critical in understanding how this organism has adapted to two hosts.

Recently we have developed the 'oligo-capping' method to produce full-length enriched cDNA libraries (3,4). Briefly, it first replaces the cap structure at the 5'-end of eukaryotic mRNA with a synthetic RNA oligo primer, then uses an oligo-dT primer for production of cDNA with reverse transcriptase. After amplification by PCR using 5'- and 3'-end primers, cDNAs are digested with restriction enzymes and ligated to a vector in a directional manner. Though a huge clone has thereby been cloned from human tissue (4), large transcripts may be under-represented in the library, and 3'-end truncated clones may be included.

As our first attempt to produce a full-length cDNA library from *P.falciparum*, we focused on the erythrocytic stage parasites, which cause lethal symptoms. *Plasmodium falciparum* 3D7 strain, which was used for the genome sequencing project, was obtained from Dr Walliker, propagated *in vitro* as described previously (5) and used for the production of the library.

DATABASE DESCRIPTION

One-pass readings of the 5'-end nucleotide sequences of 3500 randomly picked clones were determined and analyzed using BLASTN. Two thousand four hundred and ninety reads constitute the first version FULL-malaria database. As shown in Table 1, it is possible to perform the BLAST search of the database with nucleotide sequences or by gene names. The whole sequence in the database can be retrieved using ftp. Full-length clone name lists will expand as the genome sequencing proceeds.

The summary of the characterization of the database is shown in Table 2. Three hundred and ninety sequences (16%) correspond to defined *P.falciparum* genes in GenBank. They represented 98 different genes, and complete protein coding sequences are included for 48 of these genes (49%). Comparison with the complete chromosome 2 sequence revealed that 143 clones are located on this chromosome. Expression of 38 genes

*To whom correspondence should be addressed. Tel: +81 3 3443 8111; Fax: +81 3 5449 5402; Email: jwatanab@manage.ims.u-tokyo.ac.jp

among these 143 clones was confirmed, for 19 (50%) of which the cDNA clones are full-length. The most important finding was the identification of three new gene candidates that had been missed by previous computational predictions using Glimmer M. Recent disputes about the difficulty of gene prediction in the malaria genome underscore the importance of comparisons of genomic DNA with cDNA (6,7).

Table 1. Contents of the database

About this database
Retrieve our sequence data ^a
BLAST search of this database ^b
<i>In silico</i> mapping to finished chromosome sequences ^c
List of full cDNA clones ^d
Statistics of this database ^e
Related sites ^f
Acknowledgments
Who we are ...

^aYou can display clone sequences in the database by annotation category, or search clones by BLAST definition. It is also possible to download the whole sequence data.

^bYou can search the database with a nucleotide sequence using BLASTN.

^cGraphical display of expression mapping to finished chromosome sequences.

^dList of the 5'-ends of full-length cDNA clones.

^eStatistics of database; clone numbers by definition and chromosome localization and full-ratio.

^fLink to related sites.

Because chromosome 2 contains one-thirtieth of the total genome, it is expected that the FULL-malaria database contains ~1000 genes and 500 full clones. Further sequencing will reveal more full-length cDNA clones, which should help with the development of new vaccines and reagents to this disease.

As the oligo-capping method is applicable for most eukaryotic pathogens, similar approaches will provide information needed about and resources to combat various diseases.

ACKNOWLEDGEMENTS

This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports and Culture of Japan and by special coordination funds for promoting science and technology (SCF) from the Science and Technology Agency (STA) of Japan.

Table 2. Results of BLASTN of FULL-malaria

Gene ^a	390	15.7%
Genome ^a	565	22.7%
EST ^b	65	2.6%
GST ^c	47	1.9%
rRNA ^a	63	2.5%
Mitochondria ^a	2	0.1%
No hit	1324	53.2%
Human	5	0.2%
Vector	30	1.2%
Total	2490	100.0%

^aBLASTN search of nr-nt database of GenBank showed homology with genes or genome of *P.falciparum* (E value <e⁻²⁰).

^bEST reported in (8).

^cGST reported in (9).

REFERENCES

- Gardner,M.J., Tettelin,H., Carucci,D.J., Cummings,L.M., Aravind,L., Koonin,E.V., Shallom,S., Mason,T., Yu,K., Fujii,C. *et al.* (1998) Chromosome 2 sequence of human malaria parasite *Plasmodium falciparum*. *Science*, **282**, 1126–1132.
- Bowman,S., Lawson,D., Basham,D., Brown,D., Chillingworth,T., Churcher,C.M., Graig,A., Davies,R.M., Devlin,K., Feltwell,T. *et al.* (1999) The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature*, **400**, 532–538.
- Maruyama,K. and Sugano,S. (1994) Oligo-capping: a simple method to replace the cap structure of eucaryotic mRNAs with oligoribonucleotides. *Gene*, **28**, 171–217.
- Suzuki,Y., Yoshitomo-Nakagawa,K., Maruyama,K., Suyama,A. and Sugano,S. (1997) Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene*, **200**, 149–156.
- Trager,W.T. and Jense,J.B. (1976) Human malaria parasites in continuous culture. *Science*, **193**, 673–675.
- Pertea,M., Salzberg,S.L. and Gardner,M.J. (2000) Finding genes in *Plasmodium falciparum*. *Nature*, **404**, 34.
- Lawson,D., Bowman,S. and Barrell,B. (2000) Reply. *Nature*, **404**, 34–35.
- Chakrabarti,D., Reddy,G.R., Dame,J.B., Almira,E.C., Laipis,P.H., Ferl,R.J., Yang,T.P., Rowe,T.C. and Schuster,S.M. (1994) Analysis of expressed sequence tags from *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **66**, 97–104.
- Reddy,G.R., Chakrabarti,D., Schuster,S.M., Ferl,R.J., Almira,E.C. and Dame,J.B. (1993) Gene sequence tags from *Plasmodium falciparum* genomic DNA fragments prepared by the genease activity of mung bean nuclease. *Proc. Natl Acad. Sci. USA*, **90**, 9867–9871.