Sequencing and Analysis of a 40 kb Region from the *Toxoplasma gondii* Genome

Siew-Fun Wai\(^1\)\(^2\), King-Hwa Ling\(^3\), Wai-Yan Yee\(^1\)\(^2\), Rahmah Mohamed\(^1\)\(^2\) and Kiew-Lian Wan\(^1\)\(^2\)

\(^1\)Malaysia Genome Institute, UKM-MTDC Smart Technology Centre, Universiti Kebangsaan Malaysia, Bangi, Selangor, \(^2\)School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor and \(^3\)Molecular Genetics Laboratory, Department of Obstetrics and Gynecology, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor.

**Abstract**

*Toxoplasma gondii* is an intracellular parasite that causes toxoplasmosis in human and animals. In this study, a cosmid clone E10-80 carrying an approximately 40 kb region from the *T. gondii* genome was subjected to random shotgun sequencing. A total of 1153 passed reads were generated and assembled into two contigs, with sizes of 24 240 bp and 13 273 bp. Gene prediction was carried out using a combination of *in silico* methods and manual curation. The consensus sequence was found to contain eight putative genes. Five of these genes can be assigned putative functions based on *in silico* evidences, i.e., asparagine-rich antigen, 200 kDa antigen, acetyl-CoA transporter, pogo transposable element and helicase.

**Keywords:** *Toxoplasma gondii*; random shotgun sequencing; cosmid clone; genome sequence assembly; gene prediction

**Introduction**

*Toxoplasma gondii* is an intracellular protozoan parasite that causes a disease known as toxoplasmosis in a broad range of hosts. *T. gondii* infection in healthy adults is typically asymptomatic, but it can be fatal in immunocompromised individuals [1,2]. Toxoplasmosis is also known to be a major problem in animals, causing abortion and neonatal mortality in sheep, goats and cattle [3]. Control of toxoplasmosis can be achieved through chemotherapy [4] and vaccination [5]. However, despite decades of research, effective control strategies remain elusive. Thus, sequencing the genome of *T. gondii* will unravel its genetic code, and provide information on numerous potential targets for control.

*T. gondii* is also recognized as a useful model system for intracellular parasitism, largely due to its unique feature of being amenable to both classical and molecular genetics [6]. Thus, the availability of genome sequences of other apicomplexan parasites such as *Plasmodium falciparum* [7] and *Cryptosporidium parvum* [8], affords an opportunity to employ comparative genomics to identify common molecules and pathways in these parasites that may contribute towards the development of more effective control strategies.

In this study, a cosmid clone containing an approximately 40 kb region from the *T. gondii* genome was chosen for sequencing using the random shotgun strategy. The generated reads were then assembled and the resulting contig sequences were subsequently analysed using a repertoire of bioinformatics tools.

**Materials and Methods**

**Construction of random sheared library**

A single colony of the *T. gondii* RH strain cosmid clone E10-80 (isolated from a cosmid library provided by Dr. David Sibley, Washington University, St. Louis, USA) was grown overnight prior to cosmid DNA extraction using the Large Construct Kit (QIAGEN, USA). Approximately 12 μg of cosmid DNA was sheared using a nebuliser (Invitrogen, USA) at 10 psi for 60 sec. The size of the sheared DNA was determined by agarose gel electrophoresis. DNA fragments of 1-5 kb were gel-excised and purified using the QIAquick™ Gel Extraction Kit (QIAGEN, USA). The DNA fragments were then dephosphorylated with calf intestine phosphorylase (1 U/μl) at 37°C for 60 min. This was followed by phenol:chloroform extraction and ethanol precipitation. The DNA fragments were ligated to the PCR® Blunt-TOPO® vector (Invitrogen, USA), and subsequently transformed into the *Escherichia coli* strain Top 10 competent cells (Invitrogen, USA).

Author for correspondence: Associate Professor Dr. Kiew-Lian Wan, School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor DE. Tel: +603-89215997 Fax: +603-89252698 E-mail: klwan@ukm.my
Sequencing reactions

Single recombinant colonies were picked and grown in 96-deep-well plates in 2X LB broth (20 g tryptone, 10 g yeast extract, 10 g NaCl) with 100 mg/ml ampicillin. Plasmid extraction was carried out using the Montage™ Plasmid Miniprep Kit (Millipore, USA). Cycle sequencing reactions were performed using the BigDye terminator chemistry (Applied Biosystems, USA) on the Peltier Thermal Cycler (MJ Research, USA). Sample analysis was subsequently carried out on the 3100 Genetic Analyzer (Applied Biosystems, USA). Sequence data were generated from double-stranded plasmid DNA templates with T3 and T7 primers.

Sequence assembly

Raw sequence data were pre-processed using PreGap4 provided in the Staden package (staden.sourceforge.net) which allows for the automated removal of ambiguous and vector sequences. Passed sequences were then assembled into contigs using the Genome Assembly Program (GAP4) [9]. The shotgun assembly mode was adopted with a minimum end-to-end exact match of 30 bp and maximum mismatch of 5%. Where possible, problematic regions consisting of low quality and single-stranded regions, and regions with base disagreements, were re-sequenced in order to generate longer sequences from existing reads either using universal or customised oligonucleotide primers. The final orientation of the contigs was determined based on read pair information, as well as flanking cosmid vector sequences.

Sequence analysis

The GC content of the entire cosmid clone sequence and overall coding sequences was calculated using Artemis [10]. The consensus sequence for the T. gondii cosmid clone E10-80 was examined for putative protein-coding open reading frames (ORFs) using a combination of gene prediction software. These include Fgenesh [11], Geneid [12], Genemark [13] and Genscan [14]. Genes predicted by these methods were imported into the Artemis annotation platform for visual inspection and manual curation.

The amino acid sequence predicted from each gene was used as a query sequence for Basic Local Alignment Search Tool (BLAST) [15] analysis against the GenBank non-redundant protein and EST databases. BLAST matches with a score of ≥ 55 and/or an E-value of ≤ 10^{-5} was considered significant and subjected to visual inspection. Each coding region was then trimmed to its start codon, stop codon and intron-exon boundaries according to the ‘gt..ag’ rule of intronic sequence. Protein features such as transmembrane domains, signal peptides and coiled coil regions were identified using TMHMM version 2.0 [16], SignalIP [17] and COILS2 [18], respectively.

Several criteria were used to define a “coding” region in the cosmid sequence. First, the coding segment must be predicted by more than one gene prediction software used in the analysis. The reliance on multiple gene prediction methods increases confidence in the prediction. Secondly, the region must be characterised by a potential ORF, which is arbitrarily defined as a sequence region that codes for an “uninterrupted” stretch of at least 50 amino acids in length. Thirdly, the predicted coding region must be significantly matched to previously characterised genes or ESTs, and/or the protein coded by the region should reveal any of the structural features such as a signal peptide, transmembrane domain or coiled coil region that are particularly useful to characterise genes of unknown function.

Results and Discussion

Generation of random reads and sequence assembly

The application of random shotgun sequencing is appropriate for large DNA molecules such as cosmid clones (~40 kb), artificial chromosomes cloned in bacteria and yeast (100-500 kb) and bacterial genomes (1-2 Mb). The usefulness of this random shotgun approach was initially demonstrated by the sequencing of the Haemophilus influenzae genome [19]. It was subsequently utilised in the sequencing of a number of other genomes including the Drosophila melanogaster [20] and the whole human [21] genomes.

In this study, we generated 1344 reads from 672 randomly selected plasmid clones, of which 1153 (86%) passed reads were obtained. These sequences represented approximately 14-fold coverage (based on an average sequencing length of 500 bp) of the whole cosmid. According to the application of Poisson distribution [22], at this sequencing fold, the probability for a base being not represented in the sequence is predicted at as low as 8.31x10^{-7}. Thus, theoretically, the generated data should represent close to 100% of the cosmid sequence. However, in this study, the GAP4 assembly resulted in two large contigs, with sizes of 24 240 bp and 13 273 bp (Figure 1). Based on the estimated insert size of 40 kb contained in the cosmid clone, the gap between the two contigs was predicted to be approximately 2500 bp in length. Various methods have been employed to close the gap but to no avail. This may be due to the presence of secondary structures that made sequencing across the gap region impossible.

General features of E10-80

Analysis of the E10-80 sequence showed that it is has a GC content of 54.5%. This shows that the region of the T. gondii sequence analysed is very GC rich compared to the genomes of P. falciparum and C. parvum which possess a GC content of 19.4% and 30.0% respectively.
Analysis of a 40kb T. gondii genome region

The content of bases A, T, C and G in the E10-80 sequence is 22.5%, 23.0%, 27.5% and 27.0% respectively. There were eight genes predicted in the E10-80 sequence with an average of one gene in every 4854 bp. This is similar with the P. falciparum genome which has one gene in every 4338 bp. The average gene length in the E10-80 sequence is 2320 bp and is similar with the average gene length in the P. falciparum genome which is 2283 bp. The coding region covers approximately 79.0% of the entire E10-80 sequence and this is much higher than the P. falciparum genome (52.6%).

The E10-80 contains eight putative genes

Eight putative genes were predicted in the E10-80 sequence. Three of the putative genes were predicted on the forward strand, whereas another five putative genes were predicted on the reverse strand (Figure 2). The annotation of gene structure is particularly complex in eukaryotes largely due to the presence of non-coding regions (introns) that interrupt protein coding segments (exons) of a gene. In the E10-80 sequence, five of the predicted genes consist of multiple exons, while the remaining three genes consist of only a single exon. The general features of the predicted genes in E10-80 is summarised in Table 2.

<table>
<thead>
<tr>
<th>Feature</th>
<th>P. falciparum</th>
<th>C. parvum</th>
<th>T. gondii cosmid E10-80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (Mbp)</td>
<td>22.9</td>
<td>9.1</td>
<td>0.04</td>
</tr>
<tr>
<td>GC content (%)</td>
<td>19.4</td>
<td>30.0</td>
<td>54.5</td>
</tr>
<tr>
<td>No. of genes</td>
<td>5268</td>
<td>3807</td>
<td>8</td>
</tr>
<tr>
<td>Mean gene length (bp)</td>
<td>2283</td>
<td>1795</td>
<td>2320</td>
</tr>
<tr>
<td>Gene density (bp per gene)</td>
<td>4338</td>
<td>2382</td>
<td>4854</td>
</tr>
<tr>
<td>Percentage of coding region</td>
<td>52.6</td>
<td>75.3</td>
<td>79.0</td>
</tr>
</tbody>
</table>

(1). The content of bases A, T, C and G in the E10-80 sequence is 22.5%, 23.0%, 27.5% and 27.0% respectively. There were eight genes predicted in the E10-80 sequence with an average of one gene in every 4854 bp. This is similar with the P. falciparum genome which has one gene in every 4338 bp. The average gene length in the E10-80 sequence is 2320 bp and is similar with the average gene length in the P. falciparum genome which is 2283 bp. The coding region covers approximately 79.0% of the entire E10-80 sequence and this is much higher than the P. falciparum genome (52.6%).

The first gene on the forward strand consists of 21 exons. It was predicted as a secretory protein based on the presence of a signal peptide. In eukaryotes, the signal peptide targets the protein to the endoplasmic reticulum for later passage to the secretory pathway. In the endoplasmic reticulum, the protein may undergo different posttranslational modifications such as glycosylation. The postulated gene was further supported by a significant match to an EST sequence. BLAST analysis showed that gene 1 has the highest match to the asparagine-rich antigen from P. falciparum 3D7 (GenBank accession number NP700973). Asparagine-rich antigen is also known as the Asparagine-Rich Protein (ARP) [23]. ARP contains tandemly ordered tetrapeptide repetitive sequences, Asn-Asn-Asn-Met, which represent natural epitopes that react with human antibodies. It also possesses a conserved U4/U6 domain which is associated with the splicing factor PRP4 that is involved in RNA processing and modification.

Genes with putative identity

The E10-80 contains eight putative genes

Eight putative genes were predicted in the E10-80 sequence. Three of the putative genes were predicted on the forward strand, whereas another five putative genes were predicted on the reverse strand (Figure 2). The annotation of gene structure is particularly complex in eukaryotes largely due to the presence of non-coding regions (introns) that interrupt protein coding segments (exons) of a gene. In the E10-80 sequence, five of the predicted genes consist of multiple exons, while the remaining three genes consist of only a single exon. The general features of the predicted genes in E10-80 is summarised in Table 2.

<table>
<thead>
<tr>
<th>Feature</th>
<th>P. falciparum</th>
<th>C. parvum</th>
<th>T. gondii cosmid E10-80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (Mbp)</td>
<td>22.9</td>
<td>9.1</td>
<td>0.04</td>
</tr>
<tr>
<td>GC content (%)</td>
<td>19.4</td>
<td>30.0</td>
<td>54.5</td>
</tr>
<tr>
<td>No. of genes</td>
<td>5268</td>
<td>3807</td>
<td>8</td>
</tr>
<tr>
<td>Mean gene length (bp)</td>
<td>2283</td>
<td>1795</td>
<td>2320</td>
</tr>
<tr>
<td>Gene density (bp per gene)</td>
<td>4338</td>
<td>2382</td>
<td>4854</td>
</tr>
<tr>
<td>Percentage of coding region</td>
<td>52.6</td>
<td>75.3</td>
<td>79.0</td>
</tr>
</tbody>
</table>

(1). The content of bases A, T, C and G in the E10-80 sequence is 22.5%, 23.0%, 27.5% and 27.0% respectively. There were eight genes predicted in the E10-80 sequence with an average of one gene in every 4854 bp. This is similar with the P. falciparum genome which has one gene in every 4338 bp. The average gene length in the E10-80 sequence is 2320 bp and is similar with the average gene length in the P. falciparum genome which is 2283 bp. The coding region covers approximately 79.0% of the entire E10-80 sequence and this is much higher than the P. falciparum genome (52.6%).
Analysis of a 40kb T. gondii genome region

Figure 2: Schematic diagram of genes in the cosmid. The upper diagram represents the forward strand, while the lower diagram represents the reverse strand. Exons are represented by black boxes while linking lines denote introns. Genes are assigned names 1 to 8. Arrow heads show the direction of the genes. Scale symbolises the length of the composite cosmid sequence in bp.

Table 2: Overview of the structure and putative function of the predicted genes in E10-80.

<table>
<thead>
<tr>
<th>Gene [strand]</th>
<th>Putative identity</th>
<th>Start - stop positions (bp)</th>
<th>Database accession number of EST match</th>
<th>No. of exon(s)</th>
<th>No. of transmembrane domain(s)</th>
<th>Signal peptide</th>
<th>Coiled-coil region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 [F]</td>
<td>Asparagine-rich antigen</td>
<td>142 - 9154</td>
<td>TC18387</td>
<td>21</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2 [F]</td>
<td>200 kDa antigen, p200</td>
<td>9537 - 11945</td>
<td>-</td>
<td>4</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3 [F]</td>
<td>Acetyl-CoA transporter</td>
<td>12331 - 17900</td>
<td>-</td>
<td>15</td>
<td>5</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>4 [R]</td>
<td>Pogo transposable element</td>
<td>21805 - 26958</td>
<td>CB024167</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
</tr>
<tr>
<td>5 [R]</td>
<td></td>
<td>27429 - 28196</td>
<td>CB303157</td>
<td>1</td>
<td>-</td>
<td>Signal anchor</td>
<td>-</td>
</tr>
<tr>
<td>6 [R]</td>
<td></td>
<td>28829 - 33894</td>
<td>CB373435</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7 [R]</td>
<td></td>
<td>34641 - 35273</td>
<td>TC12394</td>
<td>1</td>
<td>1</td>
<td>Signal anchor</td>
<td>Yes</td>
</tr>
<tr>
<td>8 [R]</td>
<td>Helicase</td>
<td>35730 - 37934</td>
<td>-</td>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*F denotes forward while R denotes reverse

The second gene on the forward strand has four exons, and is predicted to possess two transmembrane domains. It has a significant BLAST match to the 200 kDa antigen (p200) found in Babesia bigemina (GenBank accession number AAF63787). It is believed that the B. bigemina p200 is a merozoite antigen which encodes a putative cytoplasmic structural protein. The p200 consists of a long, highly charged central repeat region of an uninterrupted alpha helix, indicative of a fibrous protein. Immunostaining electron micrograph showed that the p200 was localised in the merozoite cytoplasm, suggesting the antigen as a functional protein involved in forming filament structures within the cytoskeleton organisation [24].

The third gene on the forward strand consists of 15 exons. This gene has a unique structure as it is made up of five transmembrane domains. In addition, it is also predicted to possess a signal peptide. It showed significant similarity with the acetyl-CoA transporter in Caenorhabditis elegans (GenBank accession number NP_495969) that is involved in inorganic ion transport and metabolism.

The fourth gene on the reverse strand was identified by all the gene prediction software used and has no intron. There is only a single EST sequence that matched significantly to this gene. This gene is significantly matched to the pogo transposable element from Rattus norvegicus (GenBank accession number NP_001100664), and possesses a Kruppel-associated box (KRAB) domain which is involved in protein-protein interactions. In addition, the gene also contains the conserved centromeric protein B (CENPB) domain, a putative DNA-binding domain found in the centromere protein B and transposases.

The eighth gene is made up of ten exons. It shows a high similarity to the ATP-dependent DNA-binding helicase in Encephalitozoon cuniculi (GenBank accession number NP_597242). The encoded protein exhibited the significant match with helicases from the DEAD superfamily which are involved in DNA replication, recombination and repair.
Genes without putative identity

Genes five, six and seven are on the reverse strand, and none showed any significant match to sequences in the GenBank database. The fifth gene has only one exon and does not have any significant EST matches. However, analysis with SignalP showed the presence of a signal anchor protein. Signal anchors are "uncleaved signal peptides" which has no signal peptidase recognition site, thus is not cleaved. The sixth gene has 11 exons, is a non-secretory protein and significantly matched to four peptides which has no signal peptidase recognition site, and does not have any significant EST matches. However, analysis predicts that this sequence adopts a coiled-coil conformation. Coiled-coil proteins are believed to play an important role in protein-protein interaction.

Acknowledgements

We thank David Sibley for the T. gondii cosmid library from which clone E10-80 was isolated. This work was supported by a grant (IRPA 09-02-02-002-BTK/TD/003) from the Ministry of Science, Technology and Innovation, Malaysia, and the Wellcome Trust Research Development Award in Tropical Medicine to K.L.W.

References


