[BIO30] High-throughput sequencing and analysis of chromosome one of *Eimeria tenella* (Houghton strain)

Ling King Hwa[1], Rozita Rosli[1], Mariana Nor Shamsudin[2], Rahmah Mohamed[3], Wan Kiew Lian[3]

[1]Department of Human Growth and Development, Faculty of Medicine and Health Sciences, [2]Department of Clinical Laboratory Sciences, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, 43400 Serdang, Selangor. [3]Interim Laboratory of National Institute for the Genomics and Molecular Biology, Smart Technology Centre, UKM-MTDC, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor.

Introduction

Eimeria tenella is one of the most common *Eimeria* species that causes coccidiosis. This species is also highly pathogenic and it is the most common species occurring in the field. *Eimeria tenella* is also related to other protozoan parasites placed under the same phylum such as *Plasmodium falciparum*, *Toxoplasma gondii* and *Cryptosporidium*.

The molecular karyotype of the haploid nuclear of *Eimeria tenella*, as determined by Pulsed Field Gel Electrophoresis (PFGE) comprises 60Mb DNA contained within 14 linear chromosomes that range in size between 1 and more than 7Mb (Shirley, 2000). The four smallest chromosomes of *Eimeria tenella* are between 1-2Mb and are suitable candidates for sequencing. Chromosome one is approximately 1.05Mb in size and a previous study by Shirley *et al*, (1996) showed that chromosome one bears the arprinocid resistant genes in *Eimeria tenella*.

Materials and methods

Generation of bacterial clones

Eimeria tenella (Houghton strain) chromosome one DNA library was transformed into ElectroMAX™ DH10B™ cells (Invitrogen™) through electroporation using a BioRad GenePulser® II electroporator at 1.7kv, 100Ω and 25µF (for 2-4kb small insert library) or 2.0kV, 200Ω and 25µF (for 1-2kb small insert library). Cells were revived by using SOC medium and were plated onto LB agar with 200mg/ml IPTG, 20mg/ml X-gal and 100mg/ml ampicillin prior to blue-white selection. Cells were then incubated overnight at 37°C. All single and white bacterial colonies were picked and incubated in 96-well plates containing LB broth with 100mg/ml ampicillin.

Large-scale plasmid DNA preparation

Large-scale plasmid DNA preparations were carried out using Montage™ Plasmid Miniprep$_{96}$ Kit (Millopore™). Both vacuum manifold and robotic techniques were applied. Plasmid preparation which utilizes the manifold technique was carried out according to the manufacturer's protocol (Millopore™) whereas the robotic preparation of plasmid DNA was carried out on a MWG RoboSmart-384 (MWG Biotech, Germany) platform.

DNA sequencing

Cycle sequencing reaction for each plasmid was prepared in a 5µl cocktail and placed in a MicroAmp® Optical 96-well Reaction Plate (Applied Biosystems). The cocktail consists of 0.5µl of ABI PRISM® BigDye® Terminator v3.1, 1.0µl of 5X sequencing buffer (Applied Biosystems), 1.0µl of 5pmol T7 forward universal primer (TAA TAC GAC TCA CTA TAG GG) or 1.0µl of 5pmol TBR custom-made reverse universal primer (GCC TCT TCG GAA TTC CGC CA) and 2.5µl of plasmid DNA (approximately 200-300ng of DNA). Cycle sequencing was performed using a Peltier Thermal Cycler (MJ Research) at 96°C (initial denaturation) for 2 minutes, 96°C (denaturation) for 10 minutes, 50°C (annealing) for 5 seconds and 60°C (elongation) for 4 minutes for 100 cycles. After the completion of cycle sequencing procedure, samples were precipitated immediately using the standard ethanol/sodium acetate precipitation method provided in manufacturer's protocol (Applied Biosystems).

Ten-microliter of Hi-Di™ Formamide (Applied Biosystems) were added into every sample before being subjected to 5 minutes denaturation at 95°C. Automated sequencing was carried out by using both the 16 capillaries 3100 Genetic Analyzer (Applied

Biosystems) and the 48 capillaries 3730 DNA Analyzer (Applied Biosystems) at in the Interim Laboratory for the National Institute of the Genomics and Molecular Biology, Universiti Kebangsaan Malaysia, Selangor, Malaysia.

Preprocessing and assembly of trace files

Trace files were generated in ABI format. All trace files were preprocessed using Pregap4 version 1.3 (staden.sourceforge.net) in order to mask away vector sequence, low quality regions and contaminants. Repetitive sequences were tagged using RepeatMasker (www.repeatmasker.org) and these tags were ignored during assembly.

Assembly of trace files were carried out using PHRAP assembler (www.phrap.org). Smith-Waterman algorithm was adopted throughout the assembly. GAP4 assembler (staden.sourceforge.net) was utilized during prefinishing of the chromosome one assembly. Preloaded scoring matrices in both PHRAP and GAP4 assemblers were used during assemblies.

Prefinishing of chromosome one assembly

Contigs were initially ordered by arranging adjacent contigs based on at least two spanning paired reads. Scaffolds- or contigs-spanning Bacterial Artificial Chromosome (BAC) DNA and HAPPY map (Dear *et al.*, unpublished results) generated for chromosome one of *Eimeria tenella* (Houghton strain) were used to determine the orientation of the sequences. BAC DNAs were extracted using the QIAGEN^{TM} Large-Construct kit (QIAGEN) according to the manufacturer's protocol.

BAC clones were individually digested using *Not* I restriction enzyme (Promega Corporation) in a 20µl reaction cocktail containing 1.0µg of BAC DNA, 2.0µl of Buffer Y^+ (Promega, USA), 1.0µl of *Not* I (10units/µl) restriction enzyme and deionized water. The cocktail was incubated at 37^{o}C for 14 hours before enzyme inactivation was performed at 65^{o}C for 15 minutes. Digested large DNA fragments were electrophoresed in a 1% pulsed field grade agarose gel (Bio-Rad Laboratories) with 0.5X TBE in CHEF-DR III pulsed field gel electrophoresis platform (Bio-Rad Laboratories). Electrophoresis was carried out at 5.8V/cm for 8.5 hours at 14^{o}C with pulse times ramped from 5 to 8 seconds at 120^o angles. Lambda ladder (Bio-Rad Laboratories) and Lambda DNA/*Hin*d III (Promega Corporation) were used as markers.

Physical analyses of consensus sequences

GC content was determined using Artemis version 5.0 (www.sanger.ac.uk/Artemis/). The average GC content was calculated as the total guanine and cytosine bases over the total consensus length. GC fluctuations were determined by analyzing a series of standard deviation of the mean value at 6kb, 12kb, 18kb, 24kb and 30kb window lengths. Regions exceeded 2.5 times of standard deviation were highlighted as GC bias regions in a 200bp window length GC plot.

Simple sequence repeats were determined using Tandem Repeat Finder (Benson, 1999). Tandem repeats of all the consensus sequences were determined using the alignment score of 2 for identical matches, 7 penalties for all mismatches, gaps opening and extensions. Alignment score to report a repeat was set to 50. Maximum period size of screening was set to 500bp. Single line results were generated and manually analyzed. Non-overlapping tandem repeats were determined at minimum 95% identity cutoff and maximum of 5% indel.

Gene survey

Possible coding sequences (CDSs) were predicted using the Artemis ver 5.0. Sequences between two adjacent stop codons were translated into amino acid sequence at 50 amino acid residues cutoff. Sequences were translated in all the six reading frames. All possible coding sequences were BLASTP against non-redundant protein database of GenBank, National Center for Biotechnology Information (NCBI). A significant match is defined as alignment with *E*-value less than e^{-5} (or e^{-4} whenever the query sequence is less than 100 amino acids) and the positives identity more than 30%.

Possible CDSs without any significant matches were subjected to BLASTN against clustered expressed sequence tag (EST) sequences of *Eimeria* species from The Institute for Genomic Research (www.tigr.org), dbEST of NCBI (http://www.ncbi.nlm.nih.gov/dbEST/) and The *Eimeria* ORESTES project hosted by Universidade de Sao Paulo (http://www.lbm.fmvz.usp.br/eimeria/).

Results

Pre-analysis and generation of shotgun sequences

Both 1-2kb and 2-4kb libraries were adequate for the generation of sequences for 10X coverage of chromosome one. A total number of 11,780 white colonies were picked and cultured, hence 23,808 sequences were passed the pre-analysis screening and subjected to shotgun assembly. Successful rate of overall sequence generation was 86.6%.

Assembly and pre-finishing of chromosome one sequences

A total of 61 contigs were generated from the assembly of chromosome one shotgun sequences. These contigs represent 9.29X or 86.9% coverage of chromosome one. Average depth coverage was 10.7 reads per base.

From the dataset, 56 contigs were ordered into 11 scaffolds (884,589bp) namely SC1-11. Five contigs (namely UC1-5) were unordered (27,786bp). Figure 1 shows the PFGE of *Not I* digested BAC DNAs which were utilized as scaffolding markers during ordering of chromosome one contigs. The final contigs and scaffolds arrangements suggested 60 gaps. Of all the gaps, 45 gaps are sequencing gaps whereas 15 gaps are possible physical gaps. The predicted average size of each gap is 2293.75bp.

GC content and distribution

GC composition of all the 61 contigs is 49.35% (AT content is 50.65%). Despite the equivalent GC to AT compositions ratio, these bases are not evenly distributed throughout the chromosome. Chromosome one shows that local GC compositions undergo substantial long-range excursion from its chromosome wide average of 49.35%. The GC content varies from 35% to 60%. Overall fluctuations would be low with the least different between standard deviation values at window size of 6kb (0.050%), 12kb (0.041%), 18kb (0.039%), 24kb (0.036%) and 30kb (0.033%).

GC skews were prominently observed in three long stretches of repetitive regions (Figure 2). Besides these regions, GC composition is conserved throughout both ends of the chromosome and regions between the long stretches of repetitive sequences. These features characterized chromosome one into seven distinct regions (namely A-H).

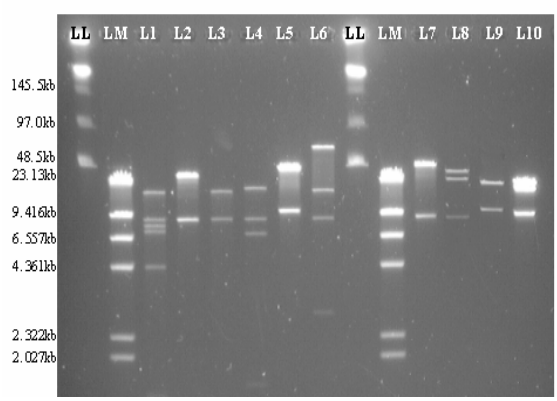


FIGURE 1 Extracted and sized BAC clones. Sized BAC clones range from 30kb to 90kb. L1 to L10 correspond to 49f09 (42.7kb), 44b11 (25.9kb), 44g07 (22.9kb), 29c03 (30.4kb), 26b05 (42.4kb), 14g03 (103.4kb), 34c04 (42.4kb), 16f01 (55.4kb), 5c09 (23.9kb) and 30c11 (43.4kb) respectively. LL is Lambda Ladder (Bio-Rad Laboratories) and LM is Lambda *Hin*d III (Promega Corporation).

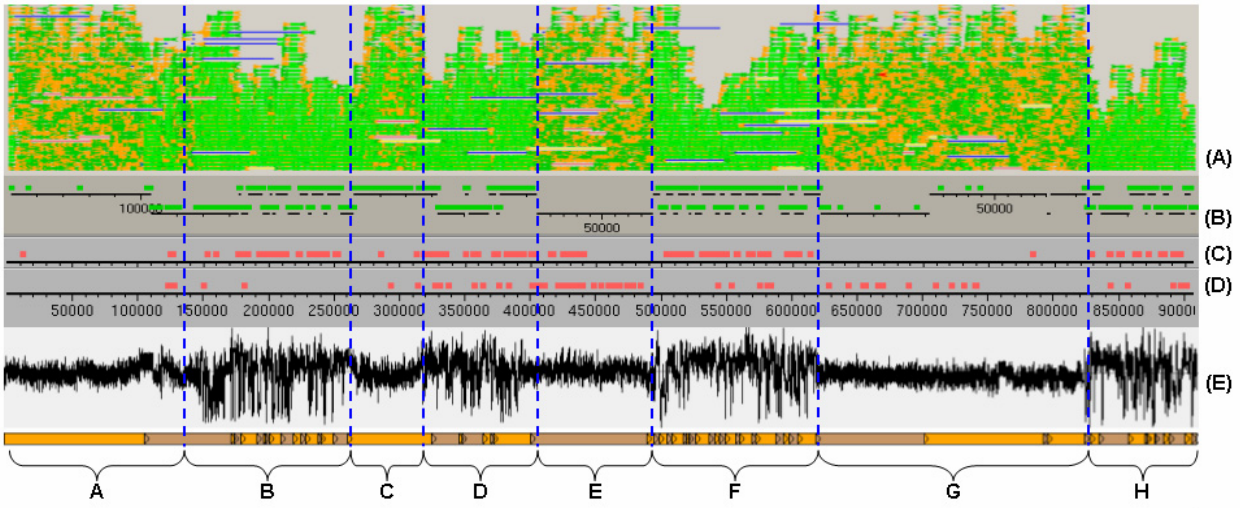
Telomeric and centromeric like regions

The telomeric like sequence TTTAGGG or its complement CCCTAAA occurs at the telomeres of chromosome one. Both ends of chromosome are flanked by multiple copies of these sequences. The proximal or 5' end and the distal or 3' end of the chromosome are characterized by 229bp and 1233bp of the exact repeat units respectively. The actual length of telomeric sequences is not known until further laboratory investigations are carried out.

On the other hand, a centromeric like region was found in chromosome one with approximately 1,453bp. This region consists of a prominent GC drop. The GC composition is as low as 18.7% compared to 81.3% of AT composition.

Simple Sequence Repeats (SSRs)

Repetitive sequences range from trinucleotide unit to minisatellite and from 2 copy numbers to more than 3,500 copy numbers. Three most frequent tandem repeats are the trinucleotide GCA and its complement TGC which occur in approximately 3,655 copy numbers followed by telomeric like sequences TTTAGGG and its complement CCCTAAA with 1,971 copy numbers and TAGC and its complement GCTA with 128 copy numbers.



Regions	Position	GC content	Repeat density$^{\xi}$	Total coding length (bp)	Coding density$^{\Psi}$
A	1…146,489	48%	9.0%	5,112	3.5%
B	146,490…273,377	49%	52.9%	22,352	17.6%
C	273,378…306,180	47%	30.6%	1,196	3.6%
D	306,181…406,783	51%	30.5%	22,469	22.3%
E	406,784…491,665	50%	13.6%	20,375	24.0%
F	491,666…620,438	53%	49.0%	36,856	28.6%
G	620,439…823,442	46%	7.4%	7,231	3.6%
H*	823,443…912375	51%	62.3%	13,109	14.7%

$^{\xi}$ Values were obtained by dividing the total length of repeats against total region length and expressed in percentage
$^{\Psi}$ Values were obtained by dividing the total coding length against total region length and expressed in percentage
* Value was obtained from unordered scaffolds and contigs

FIGURE 2 Regional characteristics of chromosome one. Position of each region was identified based on the GC plot at a global view. GC plot presented here is based on window length of 200bp. (A) is the template display of chromosome one sequences. Green tags indicate repetitive sequences. (B) shows the repeats distribution according to contigs. (C) is the region that matched with EST sequences whereas (D) is the region that matched with non-redundant database of GenBank (NCBI). Panel (E) is the GC plot of chromosome one.

Trinucleotide TGC repeats range from short to long stretches, typically from 2 copy numbers to as many as 81 copy numbers (approximately 6bp to 243bp). Telomeric like sequence repeats TTTAGGG were also found in short to long stretches, from 4 copy numbers to as many as 134 copy numbers (approximately 28bp to 938bp). TAGC was found in 7 to 21 copy numbers (approximately 28bp to 84bp) whereas TAAA occurred in 7 to 8 copy numbers (approximately 28bp to 32bp). A palindromic tandem repeat, TGCA was found to occur in 6 to 10 copy numbers (approximately 24bp to 40bp). AT rich tandem repeats, TTTTA and TTTTTA were both occurred only in 2 stretches in the draft sequence; 6 and 9 copy numbers for TTTTA (32bp and 44bp) and 5 and 6 copy numbers for TTTTTA (29bp and 36bp).

Gene content of chromosome one

From the gene survey approach, only 63 genes with known, similar or putative functions were found whereas 18 genes were unknown function or at hypothetical level. A total of 692 out of the remaining 8,705 predicted CDSs were matched significantly to EST sequences. Alignments with non-redundant databases and EST sequences give a coarse estimation of coding densities of chromosome one.

The distribution of coding sequences in chromosome one is illustrated in Figure 2. Detail observations revealed 128,700bp or 12.3% of chromosome one is coding sequences whereas 87.7% is non-coding. Chromosome one consists of a centrically dense coding sequences. Both regions A and G were found to have very low coding density values whereas regions B, D, E and F are dense with coding sequences. Average GC composition for the coding sequences was 53.9% compared to 46.1% AT composition. The GC composition of coding sequences was higher than the mean GC composition for chromosome one and also the average GC composition of non-coding sequences which is equivalent to 49.4% and 48.6% respectively.

However, all the analyses carried out were based only on homology searching. These results do not represent novel coding sequences which can only be deduced using *ab initio* annotation method. Nevertheless, the figures presented here serve as preliminary data regarding gene density in chromosome one.

Significant hits may reflect the present of the genes in chromosome one. It is also worth mentioning that a previously characterized LPMC-61 antigen which was identified from an *Eimeria tenella* sporozoite, oocysts and schizonts (Ko *et al*., 1990) was found in chromosome one at location 242,759…242938bp. The polypeptide was proposed as an important immunogen for use as a vaccine against *Eimeria tenella* since the merozoite stage of avian *Eimeria* has been implicated in the induction of a protective immune response in chickens. Other significant hits that may be worth further characterization are the circumsporozoite protein precursor-related, Elongation factor Tu, proteophosphoglycan, proteases, and AAA ATPase family proteins that are involved in the parasite's mobility, parasite-host interaction and possibly invasion.

A group of significant hits in SC11; the integrase, pol polyprotein and reverse transcriptase suggest possible LTR elements in the region. Further characterization of these genes relative to the chromosomal region will unveil the function and the main role of LTRs in chromosome one, specifically the subtelomeric and telomeric regions.

Discussion

The full shotgun sequence of the chromosome one of *Eimeria tenella* is in the process of rapidly being finished while the number of gaps minimized. This near to finished chromosome one sequence will serve as tremendous resource to all investigators studying the biology of parasites, allowing research groups to shift from hypothesis-driven ideas about proteins, DNA and RNA of *Eimeria tenella* to experimental analyses of these hypotheses. Once all the genes in the chromosome are identified, metabolic pathways can be inferred and their components, some of which will be parasite-specific, may provide potential targets for new drug design.

A common characteristic of some of the better-described protozoan genomes is the placement of genes thought to be involved in immune evasion in telomeric and subtelomeric sites on multiple chromosomes. The well characterized variant surface glycoprotein (*VSG*) loci in *Trypanosome brucei*, the *var*, *rif* and *stevor* genes in *Plasmodium*, and the trans-sialidase and mucin genes in *Trypanosome cruzi* are all in this class.

Whether the low coding density within region A and G in *Eimeria tenella* also infer the similar function is not known. Both regions A and G as well as the telomeric and subtelomeric regions in chromosome one consist of low or no homology to existing annotated proteins and ESTs. This suggests novel and non-conserved proteins that may be expressed in different stages of development or these regions are simply non-coding or that the genes located at more centromeric locations act as potential reservoirs of genetic information for the creation of additional antigenic variants for survivability. However, these previously suggested models of immune evasions have yet to be confirmed.

In *Trypanosome brucei*, three phenomena of RNA synthesis or processing are among the most distinctive features; (1) polycistronic transcription, (2) *trans*-RNA splicing and (3) kinetoplast RNA editing. Nonetheless, chromosome one of *Eimeria tenella* already demonstrates its exceptional characteristics of repeats arrangement in three distinct regions and the centrically arranged homology-based derived coding sequences. Knowledge of the overall gene organization in *Eimeria tenella* will greatly facilitate the design of experiments to elucidate the unique features of the parasite.

Chromosome one in its 'adolescence' has proven to be invaluable for the mining of genetic information in order to reveal the genomic organization of *Eimeria tenella*. As the genomic DNA sequencing enters its finishing phase, annotation, functional identification and characterization of all the predicted genes in chromosome one will be able to carry out. By applying microarrays, large-scale gene knockouts, proteomics approaches and other high-throughput methods, we will soon be able to elucidate the highly unusual cellular and molecular

mechanisms that make *Eimeria tenella* such a unique pathogen in chickens.

This very first attempt to obtain a completely finished chromosome sequence of *Eimeria tenella* will serve as model in explicating the rest of the 13 chromosomes within the genome. Complementary information from the nearly completed 8X whole genome shotgun sequences of *Eimeria tenella* is provided by Wellcome Trust Sanger Institute, Hinxton, UK. Together this information will be corroborated towards the generation of high quality consensus for better understanding of the biology of *Eimeria tenella*.

References

Shirley, M.W. (2000). The genome of Eimeria spp., with special reference to Eimeria tenella - a coccidium from the chicken. *International Journal for Parasitology* 3: 485-493.

Shirley, M.W. and Harvey, D.A. (1996). *Eimeria tenella*: genetic recombination of markers for precocious development and arprinocid resistance. *Applied Parasitology* 37: 293-299.

International Human Genome Sequencing Consortium. (2000). Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

Benson, G. (1999). Tandem repeat finder: a program to analyze DNA sequences. *Nucleic Acids Research* 27(2): 573-580.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25: 3389-3402.

Ko, C., Smith, C.K. and McDonell, M. (1990). Identification and characterization of a target antigen of a monoclonal antibody directed against Eimeria tenella merozoites. *Molecular Biochemistry Parasitology* 41(1): 53-63.