GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Sequence analysis of the *PIP5K* locus in *Eimeria maxima* provides further evidence for eimerian genome plasticity and segmental organization

B.K. Song[1], M.Z. Pan[1], Y.L. Lau[2] and K.L. Wan[3,4]

[1]School of Science, Monash University Malaysia,
Bandar Sunway, Selangor DE, Malaysia
[2]Department of Parasitology, Faculty of Medicine, University of Malaya,
Kuala Lumpur, Malaysia
[3]Malaysia Genome Institute, Kajang,
Selangor DE, Malaysia
[4]School of Biosciences and Biotechnology,
Faculty of Science and Technology, Universiti Kebangsaan Malaysia,
Bangi, Selangor DE, Malaysia

Corresponding author: B.K. Song
E-mail: song.beng.kah@monash.edu

Genet. Mol. Res. 13 (3): 5803-5814 (2014)
Received June 25, 2013
Accepted November 27, 2013
Published July 29, 2014
DOI http://dx.doi.org/10.4238/2014.July.29.8

ABSTRACT. Commercial flocks infected by *Eimeria* species parasites, including *Eimeria maxima*, have an increased risk of developing clinical or subclinical coccidiosis; an intestinal enteritis associated with increased mortality rates in poultry. Currently, infection control is largely based on chemotherapy or live vaccines; however, drug resistance is common and vaccines are relatively expensive. The development of new cost-effective intervention measures will benefit from unraveling the complex genetic mechanisms that underlie host-parasite interactions, including the identification and characterization of genes encoding

proteins such as phosphatidylinositol 4-phosphate 5-kinase (PIP5K). We previously identified a PIP5K coding sequence within the *E. maxima* genome. In this study, we analyzed two bacterial artificial chromosome clones presenting a ~145-kb *E. maxima* (Weybridge strain) genomic region spanning the *PIP5K* gene locus. Sequence analysis revealed that ~95% of the simple sequence repeats detected were located within regions comparable to the previously described feature-rich segments of the *Eimeria tenella* genome. Comparative sequence analysis with the orthologous *E. maxima* (Houghton strain) region revealed a moderate level of conserved synteny. Unique segmental organizations and telomere-like repeats were also observed in both genomes. A number of incomplete transposable elements were detected and further scrutiny of these elements in both orthologous segments revealed interesting nesting events, which may play a role in facilitating genome plasticity in *E. maxima*. The current analysis provides more detailed information about the genome organization of *E. maxima* and may help to reveal genotypic differences that are important for expression of traits related to pathogenicity and virulence.

Key words: Apicomplexan parasites; Comparative sequence analysis; *PIP5K*; Segmental organization; Coccidiosis

INTRODUCTION

Eimeria maxima is one of the seven *Eimeria* parasite species that causes avian coccidiosis, an intestinal enteritis that can lead to poor absorption of nutrients, weight loss, retarded growth rates, lowered egg production, diarrhea, and increased mortality rates in poultry (Shirley et al., 2007). The disease affects previously unexposed birds, most notably young birds, and thus incurs severe economic losses in the poultry industry. Currently, *Eimeria* spp are mainly controlled using medication with anticoccidial drugs in the feed and water or by vaccination using varied formulations of wild-type or attenuated live parasites. However, the heavy reliance on anticoccidial drugs rapidly selects for the emergence of drug-resistant populations. Unfortunately, little is known about the molecular basis of *Eimeria* drug resistance and the mechanisms involved. Similarly, the use of vaccines has limitations including the lack of cross-protection between species and among different strains of some *Eimeria* species, as well as the high cost of production and formulation (Barta et al., 1998). Research on the development of more efficient drugs and vaccines may be enhanced with additional genomic resources and in-depth comparative genomics studies of *Eimeria* spp, which could provide valuable insight into resistance mechanisms, potential targets for novel drugs and vaccine interventions, and their likely efficacy in the field.

The availability of transcriptome data (Wan et al., 1999; Amiruddin et al., 2012), linkage maps (Shirley and Harvey, 2000; Blake et al., 2011), and genomic sequences (Ling et al., 2007b; Blake et al., 2012) now provide opportunities for the identification of pathogenicity- or immunogenicity-related genes in eimerian genomes. By scrutinizing genetic codes in the genomes and unraveling the complex biochemical and genetic mechanisms that underlie

host-parasite interactions, potential targets for novel anticoccidial drugs and vaccines can be recognized (Shirley et al., 2004). Among the potential target proteins, phosphatidylinositol 4-phosphate 5-kinase (PIP5K) has previously been reported to play important roles in diverse biological processes, including signal transduction, cell secretion, vesicular trafficking, and regulation of cytoskeleton assembly (Kunz et al., 2000). For the *Eimeria* genus, PIP5K coding regions have been described for *Eimeria tenella* (Ling et al., 2007a) and more recently for *E. maxima* (Goh et al., 2011), including analysis of the *E. maxima* (Weybridge strain) *PIP5K* gene structure, comparison with related apicomplexans, and structural variations that have occurred during the evolution of the locus.

Here, we report results of a comparative genomic analysis for the characterization of *E. maxima* (Weybridge strain; hereafter designated as EmW) genomic regions around the *PIP5K* gene locus. We identified two EmW bacterial artificial chromosome (BAC) clones and determined a 145-kb nucleotide sequence encompassing the putative *PIP5K* gene. By comparing the sequence with the orthologous sequence from *E. maxima* (Houghton strain; hereafter designated as EmH), we revealed the structural organization and microcollinearity of the orthologous regions in both strains. In addition, we report the identification of segmental organization in both *Eimeria* genomes and compared it with those of similar findings as revealed in the *E. tenella* chromosome 1 sequence (Ling et al., 2007b). These results serve as a first step in providing insight into the genomic structure, gene content, biological role, and adaptive evolution of the *PIP5K* region in *Eimeria*.

MATERIAL AND METHODS

BAC clone selection and fingerprint mapping

A 490-bp probe derived from the *E. tenella PIP5K* genomic region was used to screen an EmW BAC library that was constructed in the pBACe3.6 vector, as described previously (Goh et al., 2011). Three BACs with homology to the probe, namely 1P2, 7C17, and 7K21, were identified. By excluding the redundant clone 7C17, clones 1P2 and 7K21 were selected for DNA extraction and small insert library construction. After polymerase chain reaction (PCR) verification using the primers described by Goh et al. (2011), these clones were restriction enzyme digested and mapped using *Bam*HI, *Eco*RI, and *Hin*dIII, and fractionated by electrophoresis on 1% agarose gel with 1X TAE buffer at 10 V for 20 h. The restriction fragment patterns were analyzed by generating a digital fingerprint map to establish the extent of overlapping BAC clones. Common fragments observed on the agarose gel were recorded, and detailed restriction maps of BACs 1P2 and 7K21 were constructed.

Subclone library construction, sequencing, and assembly

BAC DNA from the clones 1P2 and 7K21 were isolated and purified to generate a random shotgun library using the TOPO® Shotgun Subcloning Library Kit (Invitrogen). In addition, three libraries based on restriction enzyme digestions (*Eco*RI, *Bam*HI, or *Hin*dIII) were developed to provide an assembly framework for sequencing. Based on Sanger sequencing technology, recombinant clones were sequenced in both directions using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) and resolved in the ABI3730XL

sequencer (Applied Biosystems). Following assembly, gaps were filled by a combination of primer walking and directed PCR amplification. The draft sequences were assembled by Sequencher (v4.8; Gene Codes Corporation) to generate sequence contigs. Results with assembly ambiguities were improved by sequencing PCR amplicons spanning the regions or by re-sequencing the original clones. The error rate for the BAC clone sequences was estimated to be lower than 1 per 10,000 bp. The finished assembly was confirmed by restriction mapping, whereby observed band sizes for all digests were compared with the estimated band sizes from the assembled BAC clone sequence.

Sequence annotation and comparative analysis

Both Ab Initio and homology-based approaches were applied to identify protein-coding genes. For Ab Initio prediction, AUGUSTUS version 2.0.3 (Stanke et al., 2008) (http://augustus.gobics.de/), FGENESH+ (http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind), and GENSCAN version 1.0 (Burge and Karlin, 1997) (http://genes.mit.edu/GENSCAN.html) were used, whereas for homology-based prediction, BLASTP and BLASTX (Altschul et al., 1997) were used for comparison against the National Center for Biotechnology Information (NCBI) non-redundant protein sequence database. Manual curation was performed using Artemis (Rutherford et al., 2000). Interspersed repeats, low complexity regions, and genomic copies of putative transposable elements (TEs) were searched using RepeatMasker (http://repeatmasker.org), CENSOR (Jurka et al., 1996), and Tandem Repeat Finder (Benson, 1999). Additionally, a collinear sequence of EmH was assembled using 22 contigs retrieved from the *Emax*DB database (Blake et al., 2012), and subsequently annotated and compared with the EmW sequence. The Dotter program (Sonnhammer and Durbin, 1995) and Artemis Comparison Tool (Carver et al., 2005) were used to indicate the regions of conservation and uniqueness. The conserved regions were annotated simultaneously in both orthologous sequence regions, whereas unique non-conserved regions were analyzed individually.

RESULTS

Sequencing and mapping of BACs

The insert sizes of the *E. maxima* BAC clones 1P2 and 7K21 were estimated to be approximately 114 and 101 kb, respectively. Using a combination of enzymatic and random shotgun sub-clone libraries, a total of 3635 sequence reads were obtained. Gaps were filled using a combination of primer walking and directed PCR amplification. The finished sequence represented a consensus of 145,100 bp, which spanned the *PIP5K* locus (GenBank submission ID: 1638495). Contig sequence assembly was verified by *Eco*RI, *Bam*HI, and *Hin*dIII restriction fingerprinting analysis and mapping of BAC-end sequences (data not shown).

Annotation of the *E. maxima* BACs harboring the *PIP5K* locus

A total of 20 non-TE-related open reading frames were identified in the 145-kb fin-

ished EmW sequence using a combination of the Ab Initio gene prediction software, homology searches, and manual annotation. These predictions indicated an average of one gene per 7.3 kb (Table 1). All predicted genes were confirmed by the identification of the corresponding protein, cDNA, and/or expressed sequence tags from *Eimeria* or other apicomplexan genomes (Table 2, **Table S1**). Nine of the 20 predicted genes were similar to other apicomplexan genes with known function (E-value $<10^{-10}$).

Table 1. Features of sequenced *PIP5K* region of *Eimeria maxima* (Weybridge strain; EmW) compared to its corresponding segment in *E. maxima* (Houghton strain; EmH) and the sequence of *E. tenella* chromosome 1.

	EmW (BAC clones 1P2 and 7K21)	EmH	*E. tenella* (chromosome 1)
Size (bp)	145,100	136,980	889,314
G+C content (%)	47.3	47.0	50.3
No. of predicted genes	20	19	216
Gene density (kb per gene)	7.3	7.2	4.1

Table 2. Genes annotated in the BAC clones 1P2 and 7K21 of *Eimeria maxima*.

E. maxima (Weybridge strain) (EmW)		*E. maxima* (Houghton strain) (EmH)		Score	Identity (%)	E value	Annotation
Gene ID	Size (aa)	Gene ID	Size (aa)				
EmW-E	326	EmH-E	221	298	59	$7e^{-86}$	Hypothetical protein
EmW-D	114	N/A					Hypothetical protein
EmW-C	376	EmH-C	376	776	100	0.0	Hypothetical protein
EmW-B	642	EmH-B	642	1318	100	0.0	Hypothetical protein
EmW-A	540	EmH-A	538	1071	100	0.0	Hypothetical protein
EmW01	1146	EmH01	1146	2381	100	0.0	Phosphatidylinositol 4-phosphate 5-kinase
EmW02	751	EmH02	718	1216	98	0.0	Splicing factor protein
EmW03	301	EmH03	301	614	100	0.0	Hypothetical protein
EmW04	1296	EmH04	1296	2662	100	0.0	tRNA-splicing endonuclease positive effector protein
EmW05	608	EmH05	608	1221	100	0.0	Hypothetical protein
EmW06	245	EmH06	245	502	100	$3e^{-147}$	Hypothetical protein
EmW07	1152	EmH07	1152	2354	100	0.0	Transhydrogenase
EmW08	2871	EmH08	2870	5081	100	0.0	Hypothetical protein
EmW09	1126	EmH09	1126	2321	100	0.0	Helicase
EmW10	820	EmH10	820	1687	100	0.0	Alpha-galactosidase
EmW11	1018	EmH11	545	815	52	0.0	Nucleoside-triphosphatase
EmW12	830	EmH12	760	1082	92	0.0	CTP synthase
EmW13	1990	EmH13	1545	1144	75	0.0	Hypothetical protein
EmW14	2032	EmH14	1960	1462	62	0.0	Hypothetical protein
EmW15	1000	EmH15	962	1292	100	0.0	Transporter/permease protein

Repetitive sequences

In the 145-kb sequence of the EmW *PIP5K* region, a total of 441 simple sequence repeats (SSRs) were identified and accounted for 21 and 40% of the whole *PIP5K* and the

feature-rich (R)-segment region, respectively (Table S2). This included 6 dinucleotide repeats (15 repeat units minimum), 97 trinucleotide repeats (8 repeat units minimum), 33 tetranucleotide repeats (6 repeat units minimum), 2 pentanucleotide repeats (6 repeat units minimum), 9 hexanucleotide repeats (6 repeat units minimum), and 62 copies of telomere-like repeats (AGGGTTT)n (Table S2). The remaining SSRs were mononucleotide repeats and other SSRs with consensus patterns spread over a wide size range (8 to 445). The overall SSR density was one per 329 bp. Di-, tri-, and tetra-nucleotide SSRs accounted for 1.4, 22, and 7.5% of the SSRs, respectively. Approximately 95% (422 of 442 loci) of the SSRs detected were scattered within the R-segments. Eighty-seven copies of putative or partial retrotransposons and 16 transposons were detected in the *PIP5K* region, corresponding to a masked base percentage of 19%. A total of 13 tRNAs were identified in the EmW contig, forming a poly-tRNA-like sequence.

Sequence comparison of orthologous *PIP5K* regions

Pairwise comparisons using Dotter revealed a moderately conserved genomic sequence stretch across the 145-kb EmW *PIP5K* regions (Figure 1). Nineteen of the twenty genes predicted in EmW had orthologs in EmH and exhibited high nucleotide similarities in their coding sequences, ranging from 52 to 100% similarity (Table S2). Among these predicted genes, the sizes and exon/intron structures of eight genes in EmW were found to be highly conserved with their orthologs in EmH (Figure 2). The list also included EmW01, which putatively encodes the PIP5K protein and shares 100% similarity with its orthologous counterpart in the EmH genome. On the other hand, the EmW11/EmH11 pair (nucleoside-triphosphatase) had the lowest level of similarity (52%) in the coding regions (Table 2). Five of the 20 orthologous pairs showed 100% similarity in the coding sequence (EmW01/EmH01, EmW04/EmH04, EmW07/EmH07, EmW09/EmH09, and EmW10/EmH10). Pairwise comparison of the *PIP5K* regions also revealed differences in the copy number of tRNAs, whereby 13 and 2 tRNAs were identified in arrays in the EmW and EmH genomes, respectively (Figures 1 and 2).

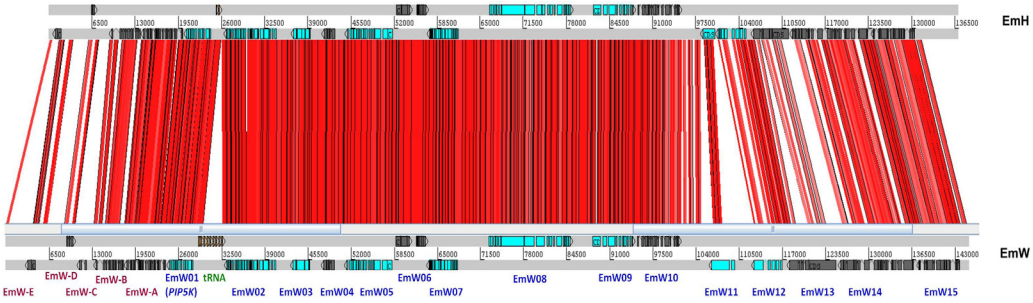

Figure 1. Collinear genomic sequence comparison of *PIP5K*-orthologous genomic regions from EmW and EmH. Red areas represent homologous regions. EmW shows high sequence similarity with the EmH counterpart, both in the genic and intergenic sequences. The difference between EmW and EmH *PIP5K* region was mainly attributed to transposable elements (TEs). Non-conserved regions (white areas) between EmW and EmH consist of a mixture of incomplete TEs and SSRs.

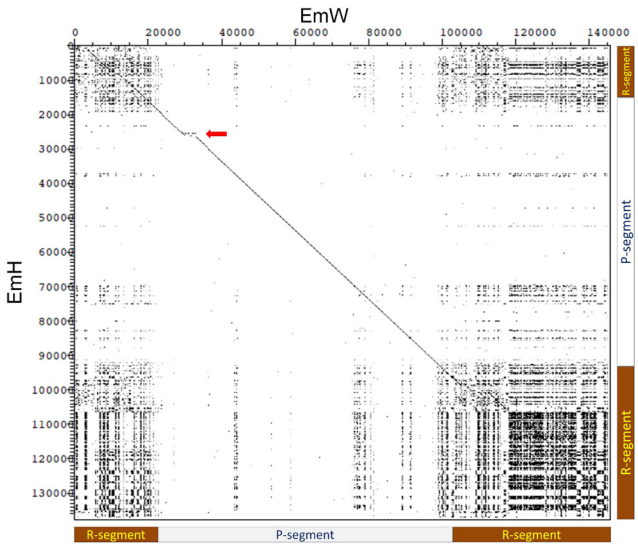

Figure 2. Comparison of the *PIP5K* region between EmW and EmH genomic sequences. Red arrow indicates the cluster of tRNAs that exists in both genomes. The dot plot used a window of 30 bp and a 70% minimum match. A bipartite feature of segmentation, comparable to the P- and R-segmental organization in *Eimeria tenella* chromosome 1 (Ling et al., 2007b), can be observed.

DISCUSSION

The EmW BAC clones 1P2 and 7K2 selected for sequencing in this study contain the *PIP5K* locus, which has been proposed to be involved in host cell invasion in apicomplexan parasites. The assembled 145-kb sequence was annotated using bioinformatic tools with manual inspection and were compared to the orthologous region of EmH. In combination with comparative sequence analysis of the *PIP5K* orthologous genomic regions, these manual inspections of assembled sequences were supported by a highly redundant finished sequence with 5-15 coverage, minimizing the risk of spurious annotation and consequential inaccurate phylogenetic inferences (Wesche et al., 2004).

Trinucleotides are the main class of SSRs in the *PIP5K* region (21.9%). Both (CAG)n and (CTG)n are the most common trinucleotide SSRs, accounting for 85.6% of all trinucleotide SSRs in the sequenced region. Almost all of these SSR loci were found to be located in R-segments of the *PIP5K* region (96.7 and 95.5% for the CAG and CTG motif types, respectively; Table 3). This is in agreement with a previous study that showed that CAG is the most frequent trinucleotide and is confined almost exclusively to the R-segments of *E. tenella* chromosome 1 (Ling et al., 2007b). In addition, in the 145-kb EmW segment, 34.9% of trinucleotide CAG repeats (5 repeat units minimum; 51 of 146 loci found on the contig; **Table S2**) were identified within exonic regions, which is in good agreement with findings for *E. tenella* chromosome 1 (Ling et al., 2007b). Although the increased number of CAG triplets resulting in long glutamine tracts have long been reported as the main causative factor leading to Huntington's disease (Hancock et al., 2001), the function of these homopolymeric amino acids in the eimerian proteome remains unclear. One potential functional role of the SSRs,

the palindromic octamer TGCATGCA, and other repetitive sequences mainly confined to the R-regions in the EmW genome could be the support of genetic recombination and hence the promotion of chromosomal rearrangement in the *PIP5K* sequence region. Recombination and shuffling of genomes have been demonstrated to produce rapid evolution of functional proteins and ultimately be beneficial to eimerian parasites. In addition, intraspecific comparative analysis of SSR variations in the present study showed that allelic diversity was mainly due to changes in the number of repeats in the SSR locus. Such sequence polymorphisms occurring in coding regions often affect gene structures, and may be helpful to understand more about the phenotypic adaptations of organisms (Li et al., 2004).

Table 3. Comparison of trinucleotide SSR repeat counts in R- and P-segments of EmW.

Motif	R-segment (0-22 and 100-145 kb)			P-segment (22-100 kb)			Total BAC sequence (0-145 kb)		
	Exonic	Non-exonic	Total	Exonic	Non-exonic	Total	Exonic	Non-exonic	Total
CAG	27	63	90	7	19	26	34	92	116
CTG	6	75	81	2	11	13	8	86	94

At the *PIP5K* locus, gene numbers and orientation were highly conserved between the two strains, EmW and EmH, although differences in the amount of repetitive sequences and nested insertion events were observed. The presence of a high copy number of TE-like elements (approximately 150 copies) and telomeric-like heptanucleotide (AGGGTTT) n repeats (247 copies) in both termini of the EmW segments, and the fact that the Dotter program revealed a highly conserved genomic sequence stretch across the intervening 78-kb regions (located between positions ~22-100 and ~15-93 kb in EmW and its orthologous EmH segments, respectively), suggest that both of these regions are derived from an R-segment-flanked region of the P-segment, which may possibly correspond to the segmental organization of *E. tenella* chromosome 1 genomic regions (Ling et al., 2007b). In addition, the Apicomplexa-specific palindromic octamer TGCATGCA has also been found to be highly abundant in the EmW R-segments (73 copies) compared to the only 14 localized in its P-segment. Similarly, it is also reasonable to speculate that the *E. maxima* genome, like its closely related species *E. tenella* (Lim et al., 2012), harbors bipartite features of segmentation in the whole genome (Blake et al., 2011). In addition, comparison between the genome of two *E. maxima* strains showed good agreement with the segmental organization notion in eimerian genomes. In particular, regions within the EmH and EmW genomes at positions 0 to 15 and 0 to 22 kb (R-segments), respectively, harbor a high copy number of AGGGTTT and other repetitive sequences.

Although the majority of the TE coding sequences was not recognizable, probably due to repeated nested insertion over evolutionary time (Smit and Riggs, 1996), partial long terminal repeats (LTRs) and motifs of these TEs were still identifiable via prediction using CENSOR and RepeatMasker. Of a total of 224 partial TEs found in the EmW *PIP5K* region, 157 and 67 were present as nested TEs in R- and P-segments, respectively. Interestingly, within the R-segment, 23, 20, and 20 copies of the incomplete TEs belong to the RTAg4 (non-LTR/R1), Ag-Jock-1 (non-LTR/Jockey), and Sm1 (non-LTR/Penelope) elements, respectively (**Table S3**). These non-LTR families of retrotransposons accounted for approximately 40% of the TEs in the *PIP5K* R-regions of both the EmW and EmH genomes. Most

of the non-LTR elements were found nested or clustered with elements of the same type. For example, there were two clusters of three and four partial elements of RTAg4, which were found tandemly deployed at positions 120,720 and 124,036 bp in the R-region of EmW. There was also a complex of four Sm1 elements located at position 142,997 bp, which further supports the idea of an insertion preference of TEs toward the same type of TEs. Although the mechanism by which tandemly clustered TEs arise in *E. maxima* is currently unclear, many studies have demonstrated that high sequence similarity between TEs of the same family type could possibly provide the preferred integration sites for other elements of the same type (Levy et al., 2010; Gao et al., 2012). Indeed, this possible explanation of the pattern of tandem array nesting is supported by the fact that many of these fragments are identical. For instance, all seven of the RTAg4 fragments (at positions 120,720 and 126,134 bp) were identified as bases derived from the region at position 1700-2280 of the RTAg4 element. Four identical TEs, representing a 120-bp fragment of the 5' end of the Sm1 non-LTR element, were found occupying position 142,894-143,614 bp in the EmW genome. The same number of Sm1 partial elements was identified at the orthologous EmH region (position 129,617-130,337 bp). Such insertion-site preference phenomena have been reported in many other eukaryotic genomes (Kaminker et al., 2002).

Further scrutiny of the nesting pattern revealed events of microcollinearity disruption between the EmW and EmH strains. Considering that the absence of the EmH sequence in gap regions may lead to misleading conclusions, we compared only the EmH contiguous segments of 65.5 kb (29,850-95,400 bp; P-region), 8.2 kb (113,550-121,720 bp; R-region), and 5.0 kb (127,350-132,350 bp; R-region), with their corresponding EmW genomic regions. The number and frequency of partial TEs detected in EmW and EmH are shown in Table 4. As expected, a lower number of genome-specific TEs were present at the P-segments in both genomes (1 copy per 32.8 kb in both sequences analyzed) when compared with R-regions (1 copy per 1.5 kb and 1 copy per 1.6 kb in EmW and EmH segments, respectively) (**Table S3**). It may be inferred that these strain-specific partial elements found in the present study were inserted into *PIP5K* regions after the split of the two strains. However, the high copy number of incomplete TEs, coupled with the fact that none of the recognized TEs in the *PIP5K* locus is structurally intact and active, suggests that ancient transposition events might have occurred at high frequencies in both genomes. Furthermore, the number of non-autonomous TEs was also relatively higher compared with that of autonomous members in *PIP5K* regions. This is indeed consistent with the view that aging class II TEs would normally show a lower number of autonomous members (Kidwell and Lisch, 2001). Although the reasons for why these insertion or nesting events are not recent and have possibly entered the "stage of senescence" (Kidwell and Lisch, 2001) remain unclear, organisms that undergo asexual reproduction like *Eimeria* are predicted to possess a higher amount of incomplete and inactive TEs (Wright and Finnegan, 2001). While an asexual reproduction mode may benefit from the use of TEs for adaptation (Gross and Williamson, 2011), virulence genes are grouped and flanked by TEs in some pathogenic bacteria (Arnold et al., 2003). Therefore, it may be reasonable to speculate that putative "pathogenic genes" like *PIP5K* could benefit from the adjacent highly packed TE genomic segments for adaptation to the host immune system.

On the other hand, the numerous strain-specific single nucleotide polymorphisms and regional rearrangements found in the R-segments of EmH compared to EmW may serve as strong evidence for the plasticity and rearrangement of the *E. maxima* genome, which may

Table 4. Comparison of number and frequency of strain-specific transposable elements (TEs) in orthologous segments of EmW and EmH *PIP5K* genomic regions.

	Strain-specific TEs					
	8.2-kb R-segment		5.0-kb R-segment		65.5-kb P-segment	
	No.	Frequency	No.	Frequency	No.	Frequency
EmW	5	1 copy per 1.6 kb	4	1 copy per 1.3 kb	2	1 copy per 32.8 kb
EmH	3	1 copy per 2.7 kb	5	1 copy per 1.0 kb	2	1 copy per 32.8 kb

facilitate the rapid evolution, survival, and/or pathogenicity of these parasites. Lorenzi et al. (2008) observed similar organizations of clustered TEs found at syntenic break points in a comparative genomics analysis of *Entamoeba* species; a group of unicellular eukaryotes that include parasitic organisms that infect humans. It is worth noting that R-segments are found at syntenic break points between EmW and EmH. This further supports the notion that R-segments serve as recombination hot spots promoting chromosomal rearrangements. Some pathogenic prokaryotes are also known for their potential in utilizing TEs for facilitating genome rearrangements that can influence the gene regulation activities of disease-associated genes (Hacker et al., 2003; Bentley et al., 2007). The accumulated evidence from previously reported findings, especially in genomes of pathogens (Hjerde et al., 2008), *Drosophila* (Lim and Simmons, 1994), yeast (Kim et al., 1998), nematodes (Stein et al., 2003), and humans (Sen et al., 2006), indirectly support the notion that TEs and repetitive DNAs locating in the R-segments may play a role in facilitating rearrangement in eimerian genomes.

The comparative sequence analysis described herein substantiates the unique segmental organization of *Eimeria* species reported previously (Ling et al., 2007b; Blake et al., 2011; Lim et al., 2012). The comparison between two *E. maxima* strains presented here is also consistent with the hypothesis of eimerian genome plasticity associated with R-segments that was identified previously using restriction fragment length polymorphism with different strains of *E. tenella* (Ling et al., 2007b). In addition, the current analysis provides more detailed information about the organization of the *E. maxima PIP5K* locus. While such a comparative analysis is informative, some of the conclusions drawn should nonetheless be interpreted with caution considering the relatively small genome region analyzed. To fully understand the dynamics of eimerian genome evolution, more representative sequence data are required in order to provide phylogenetic support based on molecular evolution. Further comparative analysis can be carried out in other parts of the *E. maxima* genome to determine the organization and microcollinearity compared to other eimerian genomes. Large-scale studies investigating the pattern of local evolution among other related eimerian taxa, including *E. tenella* and *Eimeria acervulina* (two other species that infect domestic chickens) and *Eimeria* species that parasitize other avian and mammalian hosts, will also be required. Most importantly, the soon-to-be-completed genome sequences of *E. tenella* will further facilitate whole genome analyses and provide valuable genomic resources for understanding the relationship among eimerians in the near future. Understanding genetic polymorphisms among species that infect poultry is likely to be important for the development of novel control measures based on drugs or vaccines, as it is easier to control homogenous pathogens than highly variable pathogens. In addition, future studies may also be conducted to gain insight into how eimerian genomes utilize TEs localized in the R-segments.

ACKNOWLEDGMENTS

Research supported by the Genomics and Molecular Biology Initiatives Program of the Malaysia Genome Institute, Ministry of Science, Technology and Innovation, Malaysia (Project #07-05-16-MGI-GMB10), the Monash University (Grant #PEH-SS-1-02-2010), the Universiti Kebangsaan Malaysia Research University Grant (#DIP-2012-21), and the University of Malaya High Impact Research Fund (#E000051-20001). The authors would like to acknowledge Dr. Michael Quail for construction of the *E. maxima* BAC library and Dr. Damer Blake for provision of the BAC clones selected. We thank members of the Schaal and Olsen laboratories of Washington University in St. Louis for comments on the manuscript.

Supplementary material

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Amiruddin N, Lee XW, Blake DP, Suzuki Y, et al. (2012). Characterisation of full-length cDNA sequences provides insights into the *Eimeria tenella* transcriptome. *BMC Genomics* 13: 21.

Arnold DL, Pitman A and Jackson RW (2003). Pathogenicity and other genomic islands in plant pathogenic bacteria. *Mol. Plant Pathol.* 4: 407-420.

Barta JR, Coles BA, Schito ML, Fernando MA, et al. (1998). Analysis of infraspecific variation among five strains of *Eimeria maxima* from North America. *Int. J. Parasitol.* 28: 485-492.

Benson G (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27: 573-580.

Bentley SD, Vernikos GS, Snyder LA, Churcher C, et al. (2007). Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *PLoS Genet.* 3: e23.

Blake DP, Oakes R and Smith AL (2011). A genetic linkage map for the apicomplexan protozoan parasite *Eimeria maxima* and comparison with *Eimeria tenella*. *Int. J. Parasitol.* 41: 263-270.

Blake DP, Alias H, Billington KJ, Clark EL, et al. (2012). EmaxDB: Availability of a first draft genome sequence for the apicomplexan *Eimeria maxima*. *Mol. Biochem. Parasitol.* 184: 48-51.

Burge C and Karlin S (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268: 78-94.

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, et al. (2005). ACT: the Artemis Comparison Tool. *Bioinformatics* 21: 3422-3423.

Gao C, Xiao M, Ren X, Hayward A, et al. (2012). Characterization and functional annotation of nested transposable elements in eukaryotic genomes. *Genomics* 100: 222-230.

Goh MY, Pan MZ, Blake DP, Wan KL, et al. (2011). *Eimeria maxima* phosphatidylinositol 4-phosphate 5-kinase: locus sequencing, characterization, and cross-phylum comparison. *Parasitol. Res.* 108: 611-620.

Gross SM and Williamson VM (2011). Tm1: a mutator/foldback transposable element family in root-knot nematodes. *PLoS One* 6: e24534.

Hacker J, Hentschel U and Dobrindt U (2003). Prokaryotic chromosomes and disease. *Science* 301: 790-793.

Hancock JM, Worthey EA and Santibañez-Koref MF (2001). A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Mol. Biol. Evol.* 18: 1014-1023.

Hjerde E, Lorentzen MS, Holden MT, Seeger K, et al. (2008). The genome sequence of the fish pathogen *Aliivibrio salmonicida* strain LFI1238 shows extensive evidence of gene decay. *BMC Genomics* 9: 616.

Jurka J, Klonowski P, Dagman V and Pelton P (1996). CENSOR-a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* 20: 119-121.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, et al. (2002). The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.* 3: RESEARCH0084.

Kidwell MG and Lisch DR (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* 55: 1-24.

Kim JM, Vanguri S, Boeke JD, Gabriel A, et al. (1998). Transposable elements and genome organization: a comprehensive

survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* 8: 464-478.

Kunz J, Wilson MP, Kisseleva M, Hurley JH, et al. (2000). The activation loop of phosphatidylinositol phosphate kinases determines signaling specificity. *Mol. Cell* 5: 1-11.

Levy A, Schwartz S and Ast G (2010). Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements. *Nucleic Acids Res.* 38: 1515-1530.

Li C, Zhang Y, Ying K, Liang X, et al. (2004). Sequence variations of simple sequence repeats on chromosome-4 in two subspecies of the Asian cultivated rice. *Theor. Appl. Genet.* 108: 392-400.

Lim JK and Simmons MJ (1994). Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. *Bioessays* 16: 269-275.

Lim LS, Tay YL, Alias H, Wan KL, et al. (2012). Insights into the genome structure and copy-number variation of *Eimeria tenella*. *BMC Genomics* 13: 389.

Ling KH, Loo SS, Rosli R, Shamsudin MN, et al. (2007a). *In silico* identification and characterization of a putative phosphatidylinositol 4-phosphate 5-kinase (PIP5K) gene in *Eimeria tenella*. *In Silico Biol.* 7: 115-121.

Ling KH, Rajandream MA, Rivailler P, Ivens A, et al. (2007b). Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organization. *Genome Res.* 17: 311-319.

Lorenzi H, Thiagarajan M, Haas B, Wortman J, et al. (2008). Genome wide survey, discovery and evolution of repetitive elements in three *Entamoeba* species. *BMC Genomics* 9: 595.

Rutherford K, Parkhill J, Crook J, Horsnell T, et al. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* 16: 944-945.

Sen SK, Han K, Wang J, Lee J, et al. (2006). Human genomic deletions mediated by recombination between *Alu* elements. *Am. J. Hum. Genet.* 79: 41-53.

Shirley MW and Harvey DA (2000). A genetic linkage map of the apicomplexan protozoan parasite *Eimeria tenella. Genome Res.* 10: 1587-1593.

Shirley MW, Smith AL and Blake DP (2007). Challenges in the successful control of the avian coccidia. *Vaccine* 25: 5540-5547.

Shirley MW, Blake D, White SE, Sheriff R, et al. (2004). Integrating genetics and genomics to identify new leads for the control of *Eimeria* spp. *Parasitology* 128 (Suppl. 1): S33-S42.

Smit AF and Riggs AD (1996). Tiggers and DNA transposon fossils in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 93: 1443-1448.

Sonnhammer EL and Durbin R (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167: GC1-10.

Stanke M, Diekhans M, Baertsch R and Haussler D (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* 24: 637-644.

Stein LD, Bao Z, Blasiar D, Blumenthal T, et al. (2003). The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* 1: E45.

Wan KL, Chong SP, Ng ST, Shirley MW, et al. (1999). A survey of genes in *Eimeria tenella* merozoites by EST sequencing. *Int. J. Parasitol.* 29: 1885-1892.

Wesche PL, Gaffney DJ and Keightley PD (2004). DNA sequence error rates in Genbank records estimated using the mouse genome as a reference. *DNA Seq.* 15: 362-364.

Wright S and Finnegan D (2001). Genome evolution: sex and the transposable element. *Curr. Biol.* 11: R296-R299.