SUPPLEMENTARY INFORMATION FOR:

Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organisation

Ling, K.-H. [1, 2, *] , Rajandream, M.-A. [3, *], Rivailler, P. [4, *, ‡], Ivens, A. [3], Yap S.–J. [1, 5], Madeira, A. M. B. N. [6], Mungall, K. [3], Billington, K. [4], Yee, W.-Y. [1, 5], Bankier, A.T. [7], Carroll, F. [4], Durham, A. M. [8] , Peters, N.[3], Loo, S.-S. [1, 5], Mat-Isa, M. N. [1], Novaes, J. [6], Quail, M. [3], Rosli, R. [1, 2] , Mariana, N. S. [1, 9], Sobreira, T. J. P. [6], Tivey, A. [3], Wai, S.-F. [1, 5], White, S. [4], Wu, X. [4], Kerhornou, A.[3], Blake, D.[4], Mohamed, R. [1, 5], Shirley, M. [4], Gruber, A. [6], Berriman, M. [3], Tomley, F. [4], Dear, P. H. [7, #], Wan, K.-L. [1, 5, #]

[1]Malaysia Genome Institute, UKM-MTDC Smart Technology Centre, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor DE, Malaysia

[2]Molecular Genetics Laboratory, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor DE, Malaysia

[3]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

[4]Division of Microbiology, Institute for Animal Health, Compton Laboratory, Compton, Near Newbury, Berkshire, RG20 7NN UK

[5]School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor DE, Malaysia

[6] Departamento de Parasitologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo SP, 05508-000, Brazil.

[7]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

[8]Departamento de Ciências da Computação, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo SP, 05508-000, Brazil.

[9]Department of Medical Microbiology and Parasitology, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor DE, Malaysia.

[‡] Present address: Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA.

[*]These authors contributed equally
[#]Joint corresponding authors

HAPPY map – detailed methods and results

Method

Sporozoites of *E. tenella* Houghton in elution buffer (0.4M Na_2HPO_4/NaH_2PO_4 pH8.0.3M NaCl) were prepared as described (Eckert et al. 1995) and mixed with low-melting-point agarose in elution buffer at 41°C to give 10^6 sporozoites/ml and 1% w/v agarose. This mixture was drawn into 100μl glass micropipettes (Blaubrand; internal diameter 1.2mm) and left to set at 4°C for 5 minutes. The resulting agarose 'strings' were expelled into lysis buffer (1M Na_2EDTA pH9.5, 1% w/v lauroyl sarcosine, 1mg/ml proteinase-K), incubated at 50°C for 48 hours, washed three times for >5hr at 4°C in 50mM Na_2EDTA, 10mM TrisHCl pH8.0 and stored in this solution at 4°C.

Small DNA fragments were removed by placing the agarose strings in the well of a 1% agarose gel in TAE buffer in a BioRad CHEF DRII PFGE apparatus, and running at 8V/cm, 0.1s pulse time, 120° switching angle, 14°C for 4hr. Strings were then stored in 1mM Tris.HCl, 0.1mM EDTA (pH8.0) at 4°C. The mapping panel was prepared by equilibrating 1.4cm of stripped string into 0.5x PCR bufferII (Perkin Elmer), then melting this at 67°C in 38ml of the same buffer, mixing by gentle inversion, and pipetting 5μl aliquots into each of 88 wells of a 96-well microtitre plate; the remaining 8 wells received 5μl of water as a negative control. All samples (the "mapping panel") were pre-amplified in a volume of 7μl under mineral oil using primer-extension pre-amplification (Zhang et al. 1992), diluted to 200μl and stored at –80°C.

Sequences for mapping (markers) were selected mainly from the data generated by the chromosome-specific sequencing project. Initially, arbitrarily-chosen sequences from this dataset were used; later, sequences were chosen preferentially from contigs of the assembly wherever to validate the contig structure or to link contigs. Some early markers were derived from Chr1-enriched libraries made from the Wis strain of *E. tenella*. PCR primer design and mapping were essentially as described previously (Konfortov et al. 2000, Glöckner et al. 2002, Bankier et al. 2003). In outline, markers were scored on 5μl samples of the mapping panel using a two-phase hemi-nested PCR protocol, multiplexed in the first phase for between 96 and 384 markers. Co-segregation levels between all possible pairs of markers were computed, and used to determine the map. Details of all markers in the HAPPY map of the chromosome (Fig. SI1) are given in the accompanying Excel spreadsheet "Ling_Markers.xls".

Results – comparison of map and sequence.

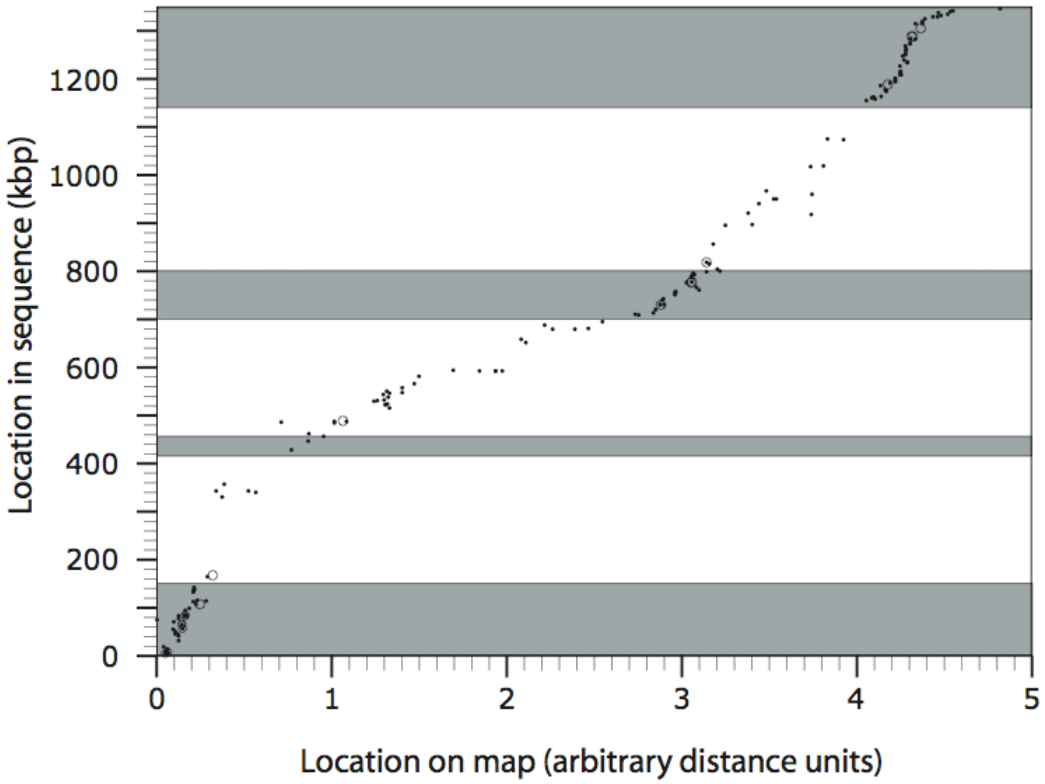

Figure SI1 – Comparison of HAPPY map and assembled sequence. The locations of markers on the HAPPY map (horizontal axis) are plotted against the locations of the corresponding sequences in the Chr1 assembly (vertical axis). Solid circles represent single-copy markers; open circles represent markers found at two or more locations in the assembled sequence (only the first occurrence is shown). The shaded rectangles correspond to the repeat-poor 'P-segments' of the sequence (see text).

The map is expanded relative to the sequence in certain regions, roughly around the centre and the quarter-points of the chromosome (Fig. SI1). Two factors may account for this. First, the mapping algorithm assumes that breakage occurs randomly along the chromosome when the DNA is sheared; but molecules on the order of a few hundred kilobases in length will tend to shear repeatedly at their mid-points, leading to an excess of breaks in the middle (and quarter-points) of the molecule and an artefactual expansion of the map in these regions. Second, the distribution of markers along the chromosome is not uniform, due largely to the abundance of repeats in some regions impeding marker selection. Fortuitously, these repeat-rich (marker-poor) regions are also at roughly the mid- and quarter-points of the chromosome, and sparsely-populated regions of linkage maps are known to expand relative to more densely populated regions.

Details of sequencing, assembly and gap closure

A total of 23,560 reads (from 11,780 randomly selected clones) were generated (with 86.6% success rate) from both libraries by sequencing the plasmid ends using the ABI PRISM® BigDye® Terminator v3.1 (Applied Biosystems) chemistry on both the 16-capillary 3100 Genetic Analyzer (Applied Biosystems) and the 48-capillary 3730 automated sequencer (Applied Biosystems).

Initial assembly of these reads were carried out using the Staden-based PHRAP (P. Green, unpublished software) which is also known as gcphrap. Default values of scoring for pairwise alignment were adopted during the banded search process. However, any pairwise alignment with less than the minimum alignment score of 30 was filtered and the alignments were discarded. Besides minimum and maximum matches, bandwidth, minimum score and penalties, stringency and completeness of assemblies were also controlled by setting additional arguments at 25 and 0.95 for –maxgap and –repeat_stringency respectively. The assembly mode with –revise_greedy and –shatter_greedy was adopted. *E*-value cutoffs were set to 1.0 and any alignment with a value higher than *E*=1.0 was discarded. In order to improve contiguity and quality of the consensus, reads from the whole-genome shotgun (WGS) sequencing project (http://www.sanger.ac.uk/Projects/E_tenella/) were incorporated into the database by using a directed method that employed Genome Assembly Program 4 (GAP4; Bonfield et al. 1995) as a platform. Additional reads from the closely-migrating chromosome 2 were also incorporated with high stringency (initial match = 300bp, mismatch = 5% and maximum pads = 25).

Contigs generated were ordered based on at least 2 consistent paired reads (with expected size range from 1 – 4kb) before subjected to BLASTN against BAC-end (ftp.sanger.ac.uk/pub/pathogens/Eimeria/tenella/BAC/) and fosmid-end (ftp.sanger.ac.uk/pub/pathogens/Eimeria/tenella/fosmid/) sequences. Relevant BAC and fosmid clones were obtained from the WTSI and the clones were sized using PFGE. BAC-end and fosmid-end sequences, and HAPPY markers were used to order the contigs into scaffolds as well as super-scaffolds. Sequencing gaps were closed by resequencing short reads (<300bp reads from bridging clones) followed by primer walking. Difficult regions were sequenced by using different combination of chemistries such as BigDye:dGTP at 3:1 ratio with/without 10% DMSO or 5M betaine as additives. For the rest of the sequence gaps, bridging clones were sonicated, end-treated and cloned by using the TA Cloning® Kit (Invitrogen) according to the manufacturer's protocol. Transposon insertion techniques were also employed as an alternative method using the EZ-Tn™ <oriV/KAN-2> Insertion Kit (Epicentre) according to the manufacturer's protocol. For gaps without any bridging clones, a high-fidelity long-ranged DNA polymerase was used to amplify the regions using specific primers with BAC, fosmid or genomic DNA as templates. Sub-libraries were also constructed from one BAC and four fosmid clones using the TOPO Shotgun Subcloning Kit (Invitrogen) according to the manufacturer's protocol.

Assembly statistics

Details of the sequence assembly are given in the table below:

Chromosome size (kb)[1]	1,050
Chr1 reads (number)[2]	16,777
Chr2 reads (number)[3]	1,336
WGS reads (number)[4]	5,072
BAC subclone reads (number)[5]	323
Fosmid subclone reads (number)[6]	1,208
Total read length (bp)	14,314,550
Expected coverage (-fold)	13.63
No. of sequence contigs	49
Total contig length (bp)	889,314
No. of sequence scaffolds[7]	9
Total scaffold length (bp)[8]	1,014,614
No. of super-scaffolds[9]	3
Total super-scaffolds length (bp)[10]	1,347,714
Physical gaps	8
Sequence gaps	40

[1] Based on pulsed-field gel electrophoresis
[2] Sequence reads from Chr1-enriched clone libraries
[3] Sequence reads from Chr2-enriched clone libraries
[4] Sequence reads incorporated from whole-genome shotgun data
[5] Sequence reads from one BAC subclone library
[6] Sequence reads from four fosmid subclone libraries
[7] Scaffolds are formed when two or more contigs are linked by at least two plasmid clone read pairs
[8] Scaffold length comprises total contig length and estimated sequence gap lengths
[9] Super-scaffolds are formed when two or more scaffolds are linked by at least one characterised BAC/fosmid clone read pair(s)
[10] Super-scaffold length comprises total contig length, sequence gap lengths and physical gap lengths; physical gaps are assumed to be as large as the maximum length of clones spanning them, and are therefore over-estimated.

Simple-sequence repeats

The table below gives a detailed breakdown of the distribution of simple-sequence repeats on the chromosome. The numbers of instances of each type of repeat (repeat unit length indicated on left, and number of tandem repeat units across top) are given for the R-fraction of the genome and, in parentheses, for the P-fraction. n.d: not determined.

Repeat unit length	**Number of repeat units**								
	3	**4**	**5**	**6-10**	**11-15**	**16-20**	**21-30**	**>30**	**Total**
2	n.d.	n.d.	n.d.	4 1	0	0	0	0	**4** **(1)**
3 **(CAG)**	n.d.	192 (6)	491 (18)	969 (6)	150	22	5	3	**1832** **(30)**
3 **(other)**	n.d.	17	15	9	0	0	1	0	**42**
4	48 (1)	113 (1)	53 (3)	82 (1)	4	1	0	0	**301** **(6)**
5	19 (1)	12	3	5	0	0	0	0	**39** **(1)**
6	15	6 (1)	4	2	0	0	0	0	**27** **(1)**
7 **(AAACCCT)**	118	186	142	108 (1)	2 (1)	1	0	0	**557** **(2)**
7 **(other)**	8	12	1	1	0	0	0	0	**22**
8-20	52	57 (1)	16	12	1	0	0	0	**138** **(1)**
>20	8	4	6	7	2	0	0	0	**27**
Total									**2989** **(42)**

The table below gives a more detailed breakdown of the distribution of tandem repeats in the repeat-rich (R) and repeat-poor (P) segments of the chromosome.

Segment	**Coordinates (bp)[1]**	**Length of sequence (bp)[2]**	**Repeat content (bp)**	**Repeat content (%)[3]**
P1	1-146782	146782	313	0.2
R1	146783-515170	170588	44073	25.8
P2	515171-557589	42419	454	1.1
R2	557590-700883	87594	12293	14.0
P3	700884-800797	99914	40	0.04
R3	800798-1143661	144264	27369	19.0
P4	1143662-1347714	197753	28	0.01
All P-segments	-	482268	835	0.17
All R-segments	-	407046	83735	20.6
Whole chr.	1-1347714	889314	84570	9.5

[1]Coordinates are given on the pseudomolecule in the EMBL file with accession number AM269894. In this file, sequence and physical gaps are in general over-estimated.
[2]Actual sequenced bases, excluding gaps.
[3]Total length of repeats as a fraction of the total length of sequenced bases.

Further details of gene prediction

Given the difficulties of gene prediction in this species, we present here a more detailed description of the annotation process and of the extent to which various predictions were supported by other data.

Overview

The 216 gene annotations were a combination of EST mapping, BLAST search and gene predictions – refined in each case by detailed inspection. Only 16 (7%) of the annotated genes were supported simultaneously by EST, BLAST and predictions; of these, only four are in the P-segments. All BLAST hits were accounted for by gene predictions and/or EST mapping. In contrast, 16 % of EST mappings (10 out of 63)were not accounted for by gene predictions or BLAST hits and this reaches (20% in the P segments). Finally, 56% of annotated genes (122 out of 216 genes) were supported only by gene prediction tools. In conclusion, gene prediction tools were the major contributors to the annotation (in some cases supported by EST hits); EST mapping added a small subset of annotations; BLAST hits alone did not contribute any further annotations, but supported many of the annotations made by other means. 66% of annotations of the P segments (84 out of 126 genes) were supported only by gene prediction tools, as compared to 43% of the R genes (38 out of 90 genes). Fig. SI2 summarizes the extent to which EST support, BLAST hits and automated predictions overlap.

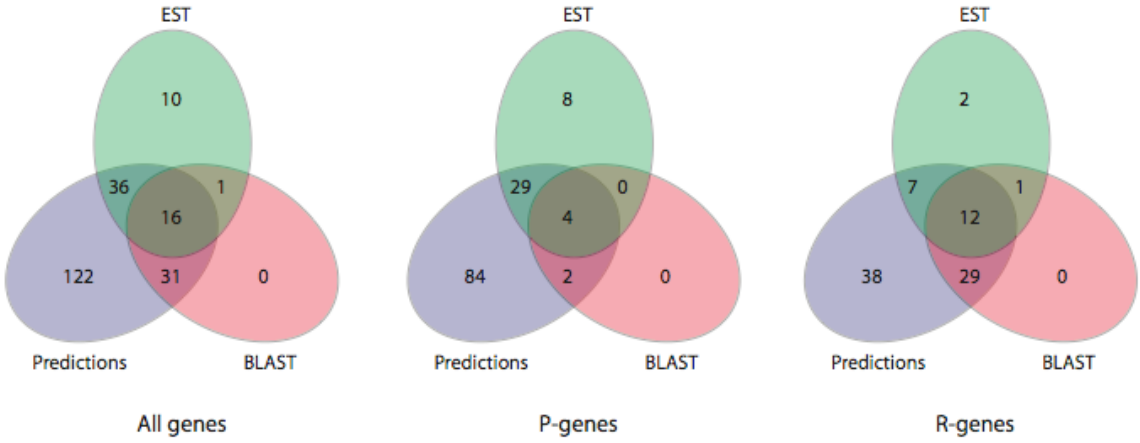

Figure S12 – concordance between evidence types for gene prediction. The Venn diagrams show (for all genes; for P-segment genes; and for R-segment genes, respectively) the number of annotations supported by various combinations of automated predictions, EST evidence and BLAST hits.

Comparison between automated gene prediction tools

Gene predictions were obtained from four tools: Glimmer HMM, SNAP, Twinscan and Genefinder. Where there were conflicts between the predictions made by each tool (often in respect of exon boundaries), predictions with similarity to other apicomplexan genes were favoured; where there was no such similarity, gene models predicted by two or more different tools were favoured. Twenty annotated genes were supported by all 4 gene prediction tools and, of these, 18 were found in the R segments. A subset of annotated genes was supported by only one prediction tool (Fig. SI3), and Glimmer HMM was the

major contributor of annotations (104 out of 205 predictions), followed by Genefinder (27 predictions), Twinscan (9 predictions) and SNAP with (4 predictions). Of the 144 predictions made by only a single tool, 74% are found in P segments and most of these (92%) are by Glimmer HMM. In the R segments, there are fewer "single tool" predictions, and most of these (23 out of 38) are by Genefinder (Fig. SI2). Each prediction tool produced some predictions that were clearly erroneous on inspection (for example, conflicting with strongly-supported predictions on the opposite strand). Glimmer and Twinscan made similar numbers (22-23) of predictions in the “wrong” strand; SNAP and Twinscan had similar numbers (45-50) of predictions that did not contribute to the gene annotations. In conclusion, all tools contributed to the annotation, but Glimmer HMM contributed the majority of predictions, particularly in the P-segments, Predictions tools agreed more in the R segments but single predictions and “wrong” predictions are found in all segments and are generated by all prediction tools.

.

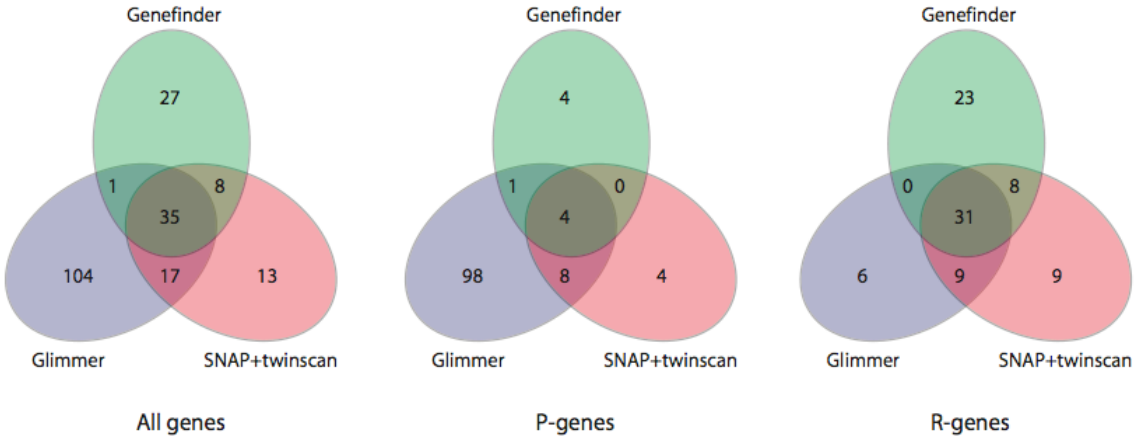

Figure SI3 – Overlap between automated predictions. The Venn diagrams show (for all genes; for P-segment genes; and for R-segment genes, respectively) the overlap between predictions made by Glimmer HMM, Genefinder and the combination of SNAPand Twinscan.

Comparison between automated gene prediction and EST/BLAST support

All four prediction tools generated similar numbers (11-12) of prediction that were supported by both BLAST hits and ESTs; also, all four gave similar numbers (19-27) of predictions supported only by BLAST hits (Fig. SI4). In contrast, 31 Glimmer predictions were supported only by ESTs, whereas only 4-7 predictions generated by the other prediction tools were supported by EST data alone. However, 56% of glimmer HMM predictions (88 out of 157 predictions) are not supported by either ESTs or BLAST hits and this proportion reaches 70% in the P segments (78 out of 111 predictions). A similar proportion was seen with Twinscan, which generated 14 predictions in the P segments of which 11 were not supported by either BLAST hits or EST data. Finally, 45% of Genefinder predictions in the R segments were not supported by either BLAST hits or EST data. 66% of ESTs mapped on to the chromosome 1 sequence were accounted for Glimmer predictions. This was in sharp contrast with the other prediction tools with 23-28% of EST data accounted for. BLAST hits were roughly equally accounted for by the prediction tools, Glimmer HMM being the best in covering BLAST hits (80%). In conclusion, even though half of the Glimmer predictions were not supported by other data, most of the mapping data (EST, BLAST) was taken into account by Glimmer HMM.

Figure SI4 – Overlap of individual gene prediction tools with EST and BLAST data. Each diagram relates to one of the four gene prediction tools (GlimmerHMM, Genefinder, Twinscan and SNAP; top to bottom) and to all annotations, to P-genes or to R-genes (left to right). In each case, the Venn diagram shows the number of annotations supported by various combinations of the prediction tool, EST hits and BLAST hits.

Analysis of exon predictions

Of the 216 annotations, 48, representing 359 exons, were supported by BLAST hits. Glimmer HMM predicted the most exons (72% of those predicted), followed by SNAP (60%), Twinscan (47%) and Genefinder (30%). SNAP generated fewest "wrong" exons, with 64% of its predicted exons contained in the final gene models. The other prediction tools are less efficient in this respect, with 55%, 37% and 25% for Glimmer HMM, Genefinder and Twinscan respectively. Finally, Glimmer HMM was the best predictor of full-length genes (13 genes), followed by Twinscan (9 genes) and Genefinder and SNAP (each 6 genes). No gene was completely predicted by all four prediction tools. Nine full genes were predicted only by Glimmer HMM whereas the other prediction tools each contributed 4-5 full genes. In conclusion, each prediction tool contributed to the annotation but no gene was fully predicted by all gene prediction tools. Glimmer HMM predicted most of the annotated exons and the Glimmer HMM prediction efficiency was comparable to, if not greater than, the prediction efficiency of the other prediction tools.

Summary

In summary, annotations were a combination of automated predictions, EST mapping and BLAST hits. Only a few annotations were supported by all three methods, and automated predictions were the major contributor to the annotation. The four automated prediction tools all contributed to the annotation, but only a small subset of genes was predicted by all four tools.

Most of the annotations of the P segments were only supported by Glimmer HMM. However, based on the efficiency of GlimmerHMM (full prediction of genes, best coverage of BLAST/EST data, average false prediction rate), we cannot dismiss these predictions.

Properties of predicted genes

Figures SI5-SI7 show the distribution of properties for predicted genes in the P and R segments of the chromosome. These distributions illustrate more precisely the general points made in the main text concerning the differences between the two populations of genes (for example, the lower incidence of introns in R-genes than in P-genes). In each case, the distributions of P- or R-genes approximate to skewed-normal distributions; a significant deviation from this (for example, a strongly bimodal distribution) would indicate the presence of a sub-population of genes of a distinct character (for example, a population of mis-predicted genes).

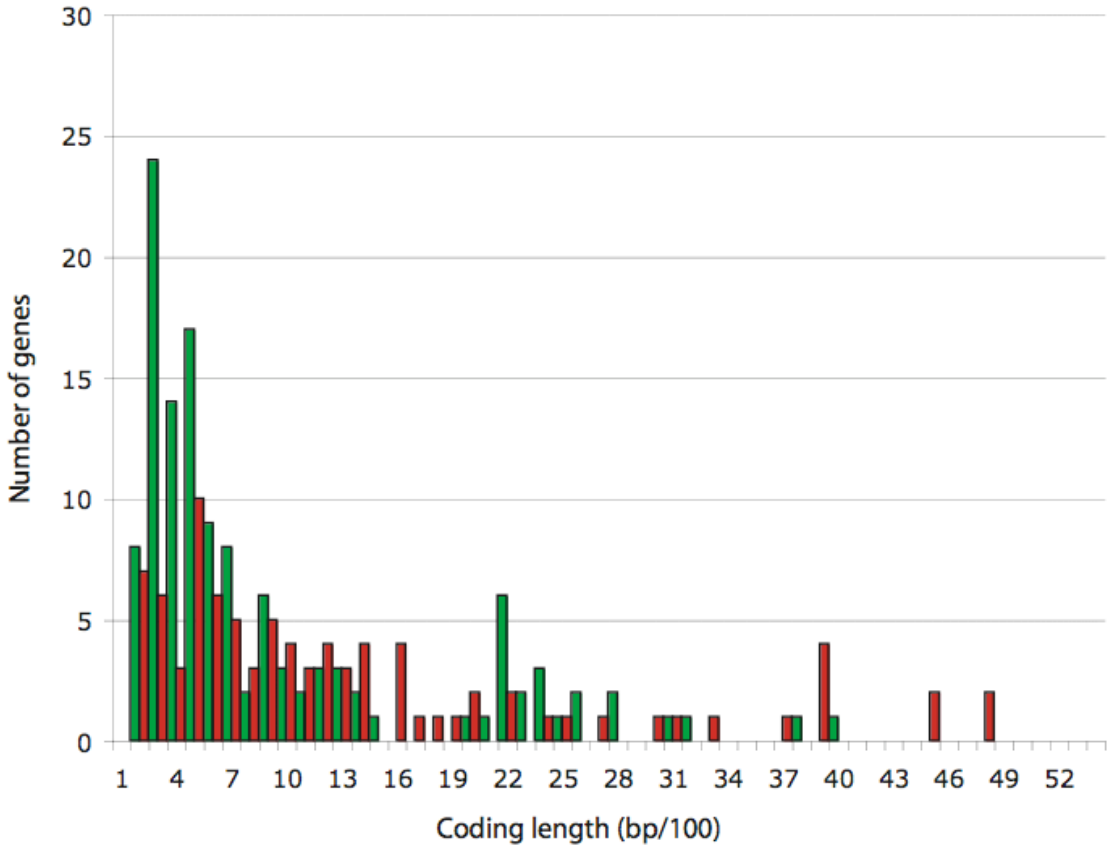

Figure SI5 – Coding-size distribution of predicted genes. The histogram shows the distribution of genes in the P-segments (green) and R-segments (red) as a function of total coding length (the combined length of all exons).

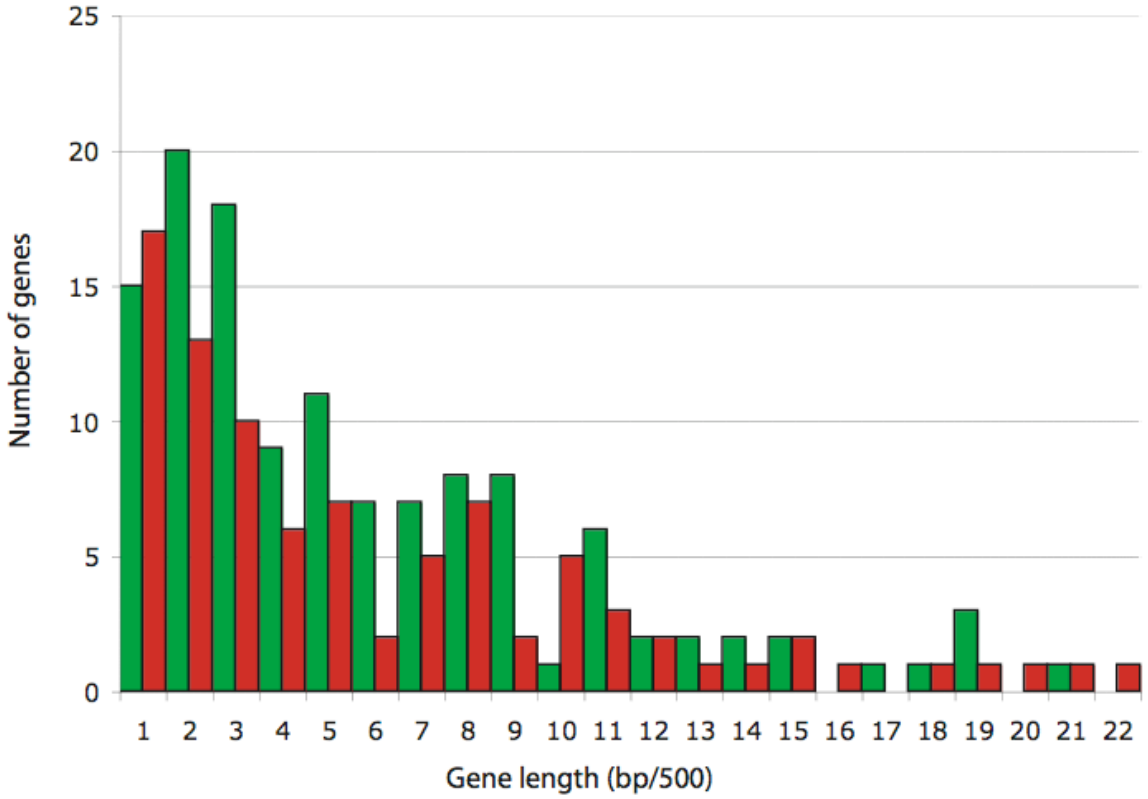

Figure SI6 – Size distribution of predicted genes. The histogram shows the distribution of genes in the P-segments (green) and R-segments (red) as a function of total gene length (the combined length of all exons and introns).

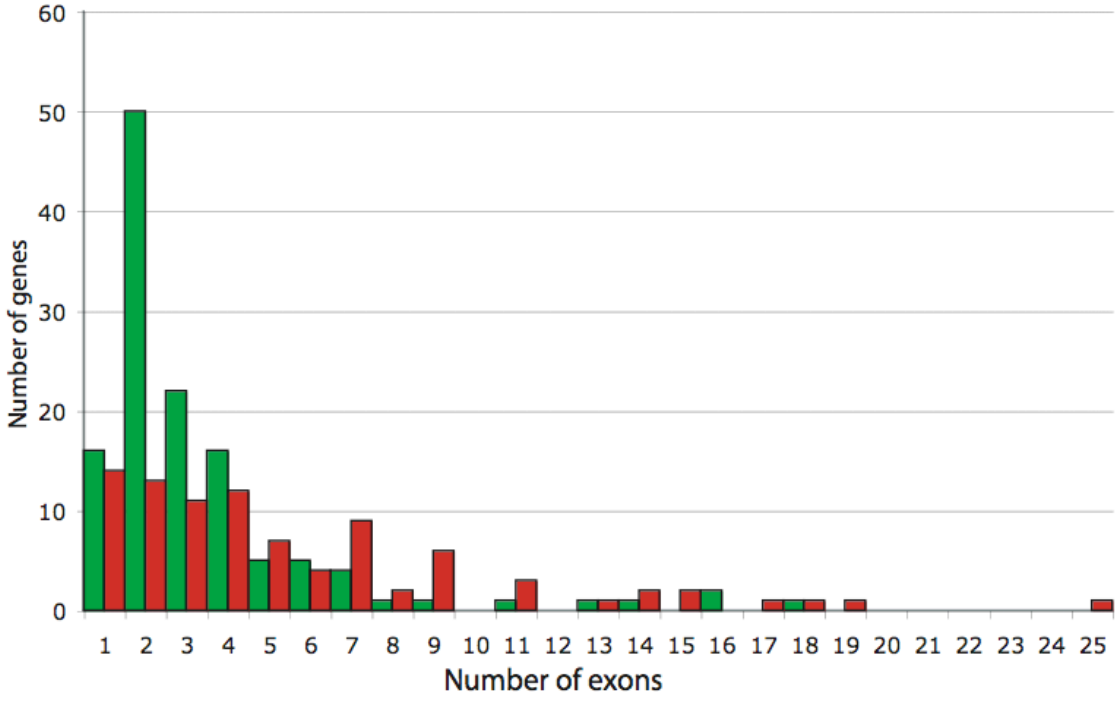

Figure SI7 - Distribution of number of exons in predicted genes. The histogram shows the distribution of genes in the P-segments (green) and R-segments (red) as a function of their number of exons.

Tandem repeats in *Eimeria* expressed sequences – comparison between species

We analysed the clustered ORESTES (Gruber, A. and Madeira, A.M.B.N., unpublished; http://www.coccidia.icb.usp.br/eimeria) from *E. acervulina*, *E. maxima* and *E. tenella*, identifying in each case the ten most abundant tandem repeat motifs, as judged by total length of repetitive sequence. Results for each species are given in the tables below.

As can be seen, $[CAG]_n$ repeats are overwhelmingly the most abundant motifs in each species; many of the other common motifs (such as $[GTTGCT]_n$) are related to this motif or its complement, or to circular permutations thereof. The telomere-like motif $[AAACCCT]_n$ is common in the ORESTES of each species; this is surprising, given its almost complete absence from the predicted transcripts of *E. tenella* chromosome 1. It is possible that chromosome 1 is atypical in this respect. The other repeats are primarily long motifs which, although not numerous, make up a disproportionate part of the total sequence due to their length.

Although there are some differences between the three species, it should be remembered that ORESTES (like all EST types) do not represent a uniform or unbiased sampling of all transcripts. Also, the data for each species is a compilation of ORESTES from several different developmental stages; the proportion of ORESTES from each stage is not the same for each species. To a first approximation, therefore, the three species exhibit similar levels and types of repeats in their expressed sequences.

Tandem repeats in *E. acervulina* clustered ORESTES:
Number of ORESTES clusters: 4,273
Total length of clusters: 3,527,292bp
Total number of repeat bases: 64,444
Repeat content: 1.83%
Ten most abundant repeats:

Total number of bases	**Period size**	**Sequence**
35132	3	CAG
3034	6	GTTGCT
2673	7	AAACCCT
2009	3	CTT
1694	9	TGCTGCTGT
1513	12	GCAACAGCAGCA
1277	4	ATGC
919	9	AGCAGCTGC
758	4	AGCT
739	3	TTG

Tandem repeats in *E. maxima* clustered ORESTES:
Number of ORESTES clusters: 3,434
Total length of clusters: 2,814,552bp
Total number of repeat bases: 37,777
Repeat content: 1.34%
Ten most abundant repeats:

Total number of bases	**Period size**	**Sequence**
18988	3	CAG
2152	6	GTTGCT
1731	9	TGCTGCTGT
1238	7	AAACCCT
837	12	TGTTGCTGCTGC
831	3	TTG
800	3	CTT
591	30	AACCATCAAGCACTGAGCCTGCAAGCACAG
562	228	TCCAGCTGCTTGCCGGCGAAAATGAGACGCT GCTGATCAGGAGGAATACCCTCCTTATCCTG AATCTTCGCCTTTACGTTTTCGATGGTGTCTG ATGGCTCAACATCCAAAGTGATGGTCTTTCC AGTCAGAGTCTTCACGAAGATCTGCATGCCA CCTCTCAAACGCAGCACCAAATGAAGGGTA GACTCCTTCTGAATGTTGTAATCGGAAAGCG TCCTGCCATCC
552	3	CCT

Tandem repeats in *E. tenella* clustered ORESTES:
Number of ORESTES clusters: 4,911
Total length of clusters: 4,042,414bp
Total number of repeat bases: 33,841
Repeat content: 0.84%
Ten most abundant repeats:

Total number of bases	Period size	Sequence
17938	3	CAG
4182	7	AAACCCT
1068	6	GTTGCT
468	6	GCTGCA
462	166	CAATCCACTGGTCCAAATGGAGGAAGAAAG CTACTTTTTCCGCATGAGCAGGTCCCAAACC CTAAACCCTAAACCCTAAACCCTAATTCGGC GTATAACCTGTTTTCCACTCAACCACCCCCA CGCACCTAGCCCAACTCCACAACTTCACAGA CCCTCTCTCCGG
441	180	CGAGCCGGACCAGACGTGCTCGTCGTCGCCC ACCAGCTCGCGGCCAAGGGAAGCCGACTAC ACACAAACCCTAAACCCTAACCACAAACCCT AAACCCTAGCAATTCCAAACCCTAAACCCCT CAAACCCTAGTCCACAGTACACTTCGCGTAA CAGCCGCCCTCTATGTTGAACACCCC
435	15	CGGTGCTCGACGCTT
413	9	GCTGCAGCA
408	147	GCAGCAGCCACCAGCTCCGCATCTCCCACGT GGAGCGCTAAACCCTAAACCCTAAACCCTAA ACCCTTTTCCACTTTCCTCCTCCTCCTCCTGC TGCTCCTGCTGCTGCTGAAGTGGCCCTCAGC GGGCTGCAGGGCTTCAAAATCA
382	27	TCCTTCTCTCTTTGCTGCTGCTGCTGT

Tandem repeats in *Eimeria tenella* expressed sequences – comparison between developmental stages

We analysed the tandem repeat content of clustered ORESTES expressed sequence tags (Gruber, A. and Madeira, A.M.B.N., unpublished; http://www.coccidia.icb.usp.br/eimeria) from various developmental stages of *Eimeria tenella*. In each case, we calculated the total length of tandemly repetitive sequence, and the percentage of the total length of clusters contributed by such sequence. Results are given in the table below. Although ORESTES do not necessarily give a completely uniform or unbiased representation of transcripts, the

data indicate a greater abundance of repeats in the transcriptomes of merozoites and sporozoites than in those of oocyst stages.

Stage	Number of ORESTES clusters	Total length of clusters (bp)	Total number of repeat bases	Repeat content (%)
Merozoite	1,267	1,012,357	9,920	0.98
Sporozoite	1,431	1,200,595	11,690	0.97
Sporulating oocysts	881	695,071	3,019	0.43
Sporulated oocysts	1,571	1,318,890	7,121	0.54
Unsporulated oocysts	860	737,996	4,713	0.64

Synteny between *Eimeria tenella* chromosome 1 and the genomes of other apicomplexans

Forty-eight of the annotated genes on *E. tenella* Chr1 encode proteins that have similarity to proteins which have been annotated or predicted in other apicomplexan genomes, with similarity to *Toxoplasma gondii* being always the top hit. Based on position of these genes within the *E. tenella* chromosome 1 assembly, six regions – containing between two and twelve genes each - can be investigated for synteny with *T. gondii* (see supplementary file "Ling_Synteny.xls" for details.)

Three instances of consecutive *E. tenella* genes having similarities to genes in the same order in the *T. gondii* genome were found, but these instances were limited to two consecutive genes in each case. Taken altogether, the 48 *Eimeria* genes have homologues in 13 different *T. gondii* chromosomes, with chromosomes Ib, VIII, X and XII having the greatest number (5 each) of homologues.. In conclusion, there is no strong evidence for conserved synteny between chromosome 1 of *E. tenella* and any part of the *Toxoplasma gondii* genome. There was insufficient homology to investigate synteny with the other apicomplexans.

Restriction-fragment length polymorphism analysis by Southern blotting – details of probes

The table below gives details of the PCR amplimers which were used as probes to detect inter-strain variations in restriction fragment lengths. "Start" and "Len" give the position of the first base of the expected amplimer in the EMBL entry AM269894 and the amplimer length, respectively.

Probe	Forward primer (5' – 3')	Reverse primer (5'-3')	Start	Len
Pa	GCACATATTAGGGCTACGTCTAGTGG	CTGTACGTGGCACAGTGTACAGGG	95403	498
Pb	GACAAAACAGTGTCGCAGAG	GTGTCCCATACCGCCACAATTC	119806	479
Pc	CTCCCGAAGTAGTAGTGCAGC	GCACCTTGTTCAGGCAACCAGCC	707361	499
Pd	CGGACCCGTTAAACTTGGCCTCGG	GCCATGACACGTCAGTCAGGGC	762770	420
Ra	CCCTAAGCCGTACTCTGCGGG	GGAGGCGACGCTGCATGC	150509	461
Rb	CGCATGCAGCGGTTTGAGC	GGGGTTGAGGTGGAGCGG	161282	499
Rc	GGGCCTTCCCTCAAAACCCTAAAC	AGCATCCAGTTAGACCCCGAAAGC	152657	400
Rd	GACTTCACACTCAGGCACATGCAA	GCAGCTGTTGGCCAGCATTAATTT	594967	227

Software availability

Custom software (PHD unpublished) is provided as supplementary files. The software is written in LabView (National Instruments), and compiled to run under MacOSX. Users will need to download and install the LabView Runtime Engine, which can be obtained free from http://digital.ni.com/softlib.nsf/websearch/9D279E79F4203562862571BF007AF5F7?opendocument&node=132070_US. For further details or for assistance, please contact PHD (phd@mrc-lmb.cam.ac.uk).

REFERENCES

Bankier, A.T., Spriggs, H.F., Fartmann, B., Konfortov, B.A., Madera, M., Vogel, C., Teichmann, S.A., Ivens, A. and Dear, P.H. 2003. Integrated mapping, chromosomal sequencing and sequence analysis of *Cryptosporidium parvum*. *Genomics* **1:** 1787-1799.

Bonfield, J.K., Smith, K.F. and Staden. R. 1995. A new DNA sequence assembly program. *Nucleic Acids Res*. **23:** 4992-4999.

Eckert, J., Braun, R., Shirley, M. and Coudert, P. (eds) 1995. Guidelines on techniques in coccidiosis research, European Commission, Brussels.

Glöckner, G., Eichinger, E., Szafranski K., Pachebat, J.A., Bankier, A.T., Dear, P.H., Lehmann, R., Baumgart, C., Parra, G., Abril, J.F. *et al*. 2002. Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature* **418:** 79-85.

Konfortov, B.A., Cohen, H.M., Bankier, A.T. and Dear, P.H. 2000. *A high-resolution HAPPY map of* Dictyostelium discoideum *chromosome 6*. *Genome Res*. **10:** 1737-1742.

Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W. and Arnheim, N. 1992 Whole genome amplification from a single cell: implications for genetic analysis. *Proc. Natl. Acad. Sci. USA* **89:** 5847-5851.