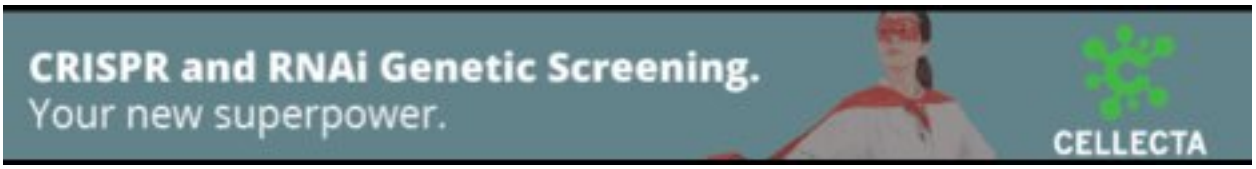
Genomic analysis of the causative agents of coccidiosis in domestic chickens

Adam James Reid, Damer Peter Blake, Hifzur Rahman Ansari, et al.

Genome Res. published online July 11, 2014
Access the most recent version at doi:10.1101/gr.168955.113



Genomic analysis of the causative agents of coccidiosis in domestic chickens

Adam J Reid[1*], Damer P Blake[2,3], Hifzur R Ansari[4], Karen Billington[3], Hilary P Browne[1], Josephine Bryant[1], Matt Dunn[1], Stacy S Hung[5], Fumiya Kawahara[6], Diego Miranda-Saavedra[7], Tareq B Malas[4], Tobias Mourier[8], Hardeep Naghra[1,9], Mridul Nair[4], Thomas D Otto[1], Neil D Rawlings[10], Pierre Rivailler[3,11], Alejandro Sanchez-Flores[12], Mandy Sanders[1], Chandra Subramaniam[3], Yea-Ling Tay[13,14], Yong Woo[4], Xikun Wu[3,15], Bart Barrell[1¥], Paul H Dear[16], Christian Doerig[17], Arthur Gruber[18], Alasdair C Ivens[19], John Parkinson[5], Marie-Adèle Rajandream[1†], Martin W Shirley[20], Kiew-Lian Wan[13,14], Matthew Berriman[1], Fiona M Tomley[2,3*], Arnab Pain[4*]

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

[2]Royal Veterinary College, Hawkshead Lane, North Mymms, Hertfordshire, AL9 7TA, UK

[3]The Pirbright Institute, Compton Laboratory, Newbury, Berkshire RG20 7NN, UK

[4]Computational Bioscience Research Center, Biological Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Jeddah, 23955-6900 Kingdom of Saudi Arabia

[5]Program in Molecular Structure and Function, Hospital for Sick Children and Departments of Biochemistry and Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

[6]Nippon Institute for Biological Science, 9-2221-1, Shin-Machi, Ome, Tokyo 198-0024, Japan

[7]Fibrosis Laboratories, Institute of Cellular Medicine, Newcastle University Medical School, Framlington Place, Newcastle upon Tyne NE2 4HH, United Kingdom

[8]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark

[9]School of Life Sciences, Centre for Biomolecular Sciences, University of Nottingham, Nottingham, NG7 2RD, UK

[10]European Bioinformatics Institute, Genome Campus, Hinxton, Cambridgeshire, CB10, 1SA

[11]Division of Viral Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

[12]Unidad Universitaria de Apoyo Bioinformático, Institute of Biotechnology, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62210, México

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

[14]Malaysia Genome Institute, Jalan Bangi, 43000 Kajang, Selangor DE, Malaysia

[15]Amgen Limited, 1 Uxbridge Business Park, Sanderson Road, Uxbridge UB8 1DH

[16]MRC Laboratory of Molecular Biology, Francis Crick Avenue, CB2 0QH, United Kingdom

[17]Department of Microbiology, Monash University, Building 76, Wellington Road, Clayton, VIC 3800, Australia.

[18]Departament of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo SP, 05508-000, Brazil

[19]Centre for Immunity, Infection and Evolution, Ashworth Laboratories, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

[20]The Pirbright Institute, Pirbright Laboratory, Ash Road, Pirbright, Surrey, GU24 0NF

† Deceased

¥ Retired

*Corresponding authors:

- Adam James Reid, Pathogen Genomics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridgeshire, CB10, 1SA. Tel: +44 (0) 01223 494810. Fax: +44 (0)1223 494919. Email: ar11@sanger.ac.uk.
- Fiona Tomley, Royal Veterinary College, Hawkshead Lane, North Mymms, Hertfordshire, AL9 7TA, UK. Tel: +44 (0) 1707 666386, fax: +44 (0) 652090, Email: ftomley@rvc.ac.uk
- Arnab Pain, Computational Bioscience Research Center, Biological Environmental Sciences and Engineering Division, King Abdullah

University of Science and Technology, Thuwal, Jeddah, 23955-6900 Kingdom of Saudi Arabia. Tel: +966 2-808-2561, Fax: n/a, Email: arnab.pain@kaust.edu.sa.

Running title: Coccidiosis genomics

Keywords: Apicomplexa, genomics, transcriptomics, host-parasite interactions, tandem repeats, protein structure, retrotransposons, surface antigens, *Eimeria*, comparative genomics, transcription factors

Abstract

Global production of chickens has trebled in the past two decades and they are now the most important source of dietary animal protein worldwide. Chickens are subject to many infectious diseases that reduce their performance and productivity. Coccidiosis, caused by apicomplexan protozoa of the genus *Eimeria*, is one of the most important poultry diseases. Understanding the biology of *Eimeria* parasites underpins development of new drugs and vaccines needed to improve global food security. We have produced annotated genome sequences of all seven species of *Eimeria* that infect domestic chickens, which reveal the full extent of previously described repeat-rich and repeat-poor regions and show that these parasites possess the most repeat-rich proteomes ever described. Furthermore, while no other apicomplexan has been found to possess retrotransposons, *Eimeria* is home to a family of chromoviruses. Analysis of *Eimeria* genes involved in basic biology and host-parasite interaction highlights adaptations to a relatively simple developmental life cycle and a complex array of co-expressed surface proteins involved in host cell binding.

Introduction

Chickens are the world's most popular food animal and the development of improved drugs and vaccines to combat poultry diseases are vital for worldwide food security. Protozoan parasites of the genus *Eimeria* cause coccidiosis, a ubiquitous intestinal disease of livestock that has major impacts on animal welfare and agro-economics. It is a particularly acute problem in poultry where infections can cause high mortality and are linked to poor performance and productivity. *Eimeria* belong to the phylum Apicomplexa, which includes thousands of parasitic protozoa such as *Plasmodium* species that cause malaria, and the widely disseminated zoonotic pathogen *Toxoplasma gondii*. *Eimeria* species have a direct oral-faecal life cycle that facilitates their rapid spread through susceptible hosts especially when these are housed at high densities (reviewed in Chapman et al. 2013). Unsurprisingly, resistance to anticoccidial drugs can evolve rapidly under these conditions and there is a continuing need to develop novel therapies (Blake et al. 2011).

More than 1200 species of *Eimeria* are described (Chapman et al. 2013) and virtually all of these are restricted to a single host species. Domestic chickens (*Gallus gallus domesticus*) can be infected by seven *Eimeria* species, each of which colonises a preferred region of the intestine causing symptoms of differing severity (Shirley et al. 2005) (Table 1). Five species induce gross pathological lesions and four of these are the most important in terms of global disease burden and economic impact (*E. acervulina*, *E. maxima*, *E. necatrix* and *E. tenella*) (Williams 1998).

Results

Genome sequences of the *Eimeria* species that infect domestic chickens

We generated annotated genome sequences of all seven species of *Eimeria* that infect domestic chickens. For the *E. tenella* a high quality reference genome incorporating annotation-directed manual improvements for targeted regions was produced for the Houghton strain (tier 1; Supplemental Table S1) as well as Illumina genomic sequencing data for the Wisconsin and Nippon strains. For *E. maxima*, *E. acervulina* and *E. necatrix* we produced draft genomes with automated post-assembly improvements (tier 2; Supplemental Table S1), and for *E. brunetti*, *E. mitis* and *E. praecox* we produced draft genomes alone (tier 3; Supplemental Table S2). The 51.8Mb *E. tenella* genome assembly corresponds well with a genomic map (described below), suggesting that it accurately reflects the true genome size (Supplemental Table S1). Furthermore the tier 1 and tier 2 assemblies are predicted to be 93-99% complete with respect to the *T. gondii* genome sequence based on the presence of core eukaryotic genes (Parra et al. 2007; Supplemental Table S1).

To investigate chromosome structures we used whole genome mapping to improve contiguity of the tier 1 and 2 genomes and were able to place up to 46% of sequence data onto fifteen or sixteen optical maps for each genome (Supplemental Table S1), which is close to the haploid chromosome number of fourteen (del Cacho et al. 2005). Although few sequence markers were available

for each chromosome we were able to map unambiguous identities to six of fifteen optical scaffolds in *E. tenella* (Supplemental Table S3).

Phylogeny and synteny between *Eimeria* species

Previous phylogenetic analyses using small numbers of sequences did not fully resolve relationships between *Eimeria* species of the chicken (Ogedengbe et al. 2011). Whole genome phylogeny shows robust separation of *E. tenella* and *E. necatrix* from the other species, as well as separation of *E. mitis* and *E. brunetti* from *E. praecox*, *E. maxima* and *E. acervulina* (Fig 1A). In support of this phylogeny there is extensive synteny between the genomes of *E. tenella* and *E. necatrix* (i.e. many orthologous genes in the same order across contigs, with only a small number of rearrangements (Fig 1B). There is notably less synteny between the genome of *E. tenella* and those of *E. maxima* and *E. acervulina* with much of the chromosome structure rearranged, although presumably retaining the same number of chromosomes. Synteny between the genomes of *E. tenella* and *Toxoplasma gondii* was non-existant, with no more than three orthologues (ETH_00031645, ETH_00031660 and ETH_00031665) found in the same order.

***Eimeria* chromosomes display a banded pattern of repeat-poor and repeat-rich regions**

Analysis of *E. tenella* chromosome 1 revealed alternating regions of repeat poor (P) and repeat rich (R) sequences (Ling et al. 2007). We now find this feature is conserved in all chromosomes of *E. tenella* and across the genomes of all *Eimeria* species examined (Fig 1B). The short tandem repeat content of each genome shows a bimodal distribution with a high frequency peak close to zero and a broad, low frequency peak with a mean around 20-30% (Supplemetal Fig S1A). This confirms a bipartite structure for the genome. Any region of the genome is either repeat rich (R; which we define as having a repeat density greater than 5%) or essentially repeat free (P). The precise repeat content differs between species with *E. tenella* having fewer R regions than other species. Of note *E. necatrix* is more repeat-rich in regions syntenic with *E. tenella* as well as across the genome generally (Supplemetal Fig S1B). Fifty-three point five percent of *E.*

tenella genes were found in repeat rich regions, suggesting that there is no preference for repeats to occur in gene-poor regions.

Eimeria protein-coding sequences are extremely rich in Homopolymeric Amino Acid Repeats (HAARs)

Short tandem repeats (STRs) can result in strings of single amino acids within predicted protein sequences. The extent of Homopolymeric Amino Acid Repeats (HAARs) is greater in *Eimeria* than in any other organism sequenced to date (Fig. 2A) and the distribution of homopolymer types is quite different from even closely related organisms such as *T. gondii* and *P. falciparum*. The most common STR in *Eimeria* species is the trinucleotide CAG (Fig 2B), which occurs preferentially in coding sequences (Fig 2C). CAG can potentially encode alanine (A), glutamine (Q), serine (S), cysteine (C) and leucine (L), but in *E. tenella* repeats are rarely translated as C or L. They are found preferentially as A, Q or S (Fig 2D). HAARs of this type, encoding strings of at least seven amino acids, occur in 57% of *E. tenella* genes, with an average of 4.3 copies per gene. We confirmed that repeats are transcribed and translated in *E. tenella*, with the data predicting similar proportions of each HAAR type to that found in the genome (Fig 2D).

An analysis of Gene Ontology terms showed that genes containing HAARs did not cluster in any particular functional class. However those involved in information processing tasks such as translation, chromatin assembly, gene expression and DNA metabolism had fewer HAARs than expected by chance (Supplemental Table S4). Indeed genes conserved across the eukaryotes had an overall lower than average repeat content (2.5% vs. 4.68% for all *E. tenella* genes). Proteins that are generally considered to be involved in host-parasite interactions such as SAGs, ROP kinases and MICs had even fewer HAARs on average than those conserved across eukaryotes (0%, 1.88% and 1.66% respectively vs. 2.5%).

We hypothesised that because HAARs are so common in *Eimeria* species, they are unlikely to interfere with protein structure and function. Indeed only 3.2% of *E. tenella* HAARs (687 of 21191) occur in Pfam domains, which make up 12.4% of *E. tenella* protein sequences. By examining conserved proteins with known 3D

structures we found that serine and glutamine HAARs tend to be insertions in loop or turn regions with medium to high solvent accessibility suggesting they do not affect protein folding (Supplemental Table S5). Alanine HAARs often align to helical regions and may result in very similar local structure (Perutz et al. 2002). Furthmore, homology modelling showed that HAARs tend to be located on the outside of proteins, away from regions involved in domain-domain interactions and active sites (Supplemental Fig. S2).

Comparative genomics of the Coccidia and wider Apicomplexa

8603 protein-coding genes are predicted in the *E. tenella* assembly, significantly more than the 7286 found in the related coccidian *T. gondii* (Supplemental Table S1), despite *T. gondii* having a nuclear genome that is about 20% (10Mb) larger. By transcriptome sequencing we identified expression of 76% of predicted *E. tenella* genes (6700) across four developmental life stages (unsporulated oocyst, sporulated oocyst, sporozoite and merozoite). The median sequence identity between *E. tenella* and *T. gondii* orthologous protein sequences was 39.7%, suggesting a large amount of sequence divergence between the two.

We identified several novel *Eimeria*-specific gene families (Supplemental Table S6; Supplemental Dataset S1). We found that two of these families (*esf1* and *esf2*) have higher K_a/K_s ratios than other genes (Supplemental Figure S3; Supplemental Dataset S2). This suggests that they may be under diversifying selection and could be important for host-parasite interactions.

The rhoptry organelles of *T. gondii* contain 30-50 kinases and pseudokinases (ROPKs; Peixoto et al. 2010; Talevich and Kannan 2013), some of which are involved in remodeling the intermediate host cell and protecting the parasite against host defenses (Saeij et al. 2007; Fentress et al. 2010). Recent analysis showed that *E. tenella* has 28 ROPK genes, including a subfamily not found in *T. gondii* (Talevich and Kannan 2013). We were able to identify orthologues of all these genes in *E. necatrix* however there is divergence in the more distantly related *Eimeria* species (Supplemental Table S7; Supplemental Dataset S1). The overall protein kinase (PK) complement of *Eimeria* species (63-84 PK) is smaller than that of *T. gondii* (128 PK) (Peixoto et al. 2010) and *Plasmodium* species (85-

99 PK) (Ward et al. 2004; Anamika et al. 2005; Miranda-Saavedra et al. 2012). This is not due solely to fewer ROPKs and FIKKs (an apicomplexan family highly expanded in *Plasmodium*) but also a reduction in CMGC kinases (the group which includes cyclin dependent kinases; Supplemental Table S7). It is proposed that CMGC kinases have evolved independently within the Apicomplexa to provide specialized functions related to lifecycle transitions (Talevich et al. 2011). Reduction of CMGC kinases in *Eimeria*, which has a simple life cycle and no intermediate host, may be an example of this specialisation.

Metabolism is well conserved between *Eimeria* and *Toxoplasma* (Supplemental Fig S4; Supplemental Dataset S3), with the clearest difference being additional enzymes involved in *Eimeria* carbohydrate metabolism. Three of these catalyse reactions in the mannitol cycle, known to be essential for survival of *Eimeria* parasites and not present in other coccidian lineages (Schmatz et al. 1989; Liberator et al. 1998).

The apicoplast is a symbiotic plastid present in most apicomplexans and known to be essential for survival of *T. gondii* (He et al. 2001). Most ancestral plastid genes have moved into the nuclear genome but many of the gene products are post-translationally imported into the apicoplast. The mechanism of import to the apicoplast is poorly understood but two proteins, Tic20 and Tic22, are thought to mediate crossing of the innermost membrane (van Dooren et al. 2008; Lim and McFadden 2010). Genes encoding Tic20 and Tic22 are not found in the *Eimeria* species studied, suggesting either a distinct mechanism for crossing the apicoplast inner membrane or a change in apicoplast function in *Eimeria*.

Apicomplexan genomes have a paucity of common eukaryotic transcription factors (Coulson et al. 2004); instead the major regulators of stage-specific gene expression genes containing DNA-binding domains of the ApiAP2 family (Balaji et al. 2005; Campbell et al. 2010). In *Eimeria* the number of genes containing ApiAP2 domains was found to vary from 44 to 54 (Supplemental Fig S5; Supplemental Dataset S1). We clustered genes containing ApiAP2 DNA-binding domains from apicomplexans and representative outgroups and identified 121

orthologous groups (Supplemental Fig. S6). We found 21 *Eimeria*-specific ApiAP2 groups, 22 further groups shared by *Eimeria* and other Coccidia and five pan-apicomplexan clusters (apiap2_og_336, apiap2_og _90, apiap2_og _1428, apiap2_og _546, apiap2_og _456; Supplemental Dataset S1).

We found a positive correlation between the number of ApiAP2 genes and genome size across the Apicomplexa ($r^2 = 0.92$; Pearson; Supplemental Fig. S5). This suggests that although *Eimeria* has a relatively simple lifecycle compared to some other genera, there is greater complexity in regulating its genome. Thus, we propose that across *Apicomplexa*, it is not the complexity of the developmental lifecycle that determines the complexity of regulation, but the amount of genome to be regulated.

Analysis of chromosome 1 of *E. tenella* identified retrotransposon-like elements (Ling et al. 2007) and we now confirm that these are related to Long Terminal Repeat (LTR) retrotransposons from the group of chromoviruses (Supplemental Fig. S7). Chromoviruses are widespread among eukaryotic genomes but have not previously been identified in apicomplexans (Kordis 2005). With the exception of *eimten1* in the *E. tenella* and *E. necatrix* genomes, the putative transposons are highly fragmented and diverged, indicating very little recent activity (Supplemental Figs S8-S9; Supplemental Table S8). Retrotransposons cannot have been transferred horizontally from the host, because the chicken genome does not contain chromoviruses (Kordis 2005). Phylogenetic analysis of predicted reverse transcriptases from *Eimeria* and other species did not robustly support a closer relationship with either plant/algal or vertebrate/fungal lineages (Supplemental Fig S7).

Eimeria-specific surface antigen genes

All apicomplexan genomes examined to date possess gene families encoding antigenic proteins that are expressed on the surface of invasive stages and thought to be important in interaction with the host immune system (Spence et al. 2013). The principal surface antigen gene family in *E. tenella* is *sag*, which

encodes single domain, membrane-bound proteins tethered by GPI anchors to the surface of invasive sporozoites and merozoites (Tabares et al. 2004).

In *E. tenella* the majority of *sag* genes are tandemly arrayed in four clusters, each on a different chromosome (Fig 3A). There are three subfamilies of *sag* genes: *sagA* is common to all species; *sagB* is restricted to *E. tenella* and *E. necatrix*; and *sagC* is restricted to the other species, being most expanded in *E. brunetti* and *E. mitis* (Table 1). All subfamilies encode signal peptides and addition sites for GPI-anchors but the *sagC* extracellular domain contains only four conserved cysteines whereas *sagA* and *sagB* have six. *SagB* and *sagC* genes each have five exons suggesting they may be more closely related to each other than to *saga* genes, which have four exons. The presence of s*agA* genes in all the *Eimeria* species suggests that these provide a core function, while *sagB* and *sagC* genes may provide functions specific to the different clades.

One core function of SAG proteins may be attachment to host cells prior to parasite invasion. In *T. gondii* the *srs* genes play a role in primary attachment and it is known that *E. tenella* SAG1 binds mammalian cells (Jahn et al. 2009). We found that multiple SAGA, but not SAGB proteins, were able to bind cultured cells (Fig 3B) suggesting that attachment is a potential function of the *sagA* family.

It is of key importance in designing vaccines to understand the array of antigens presented to the host immune system. Using single-cell RT-PCR we found that multiple members of the *sagA* and *sagB* subfamilies were co-expressed in individual sporozoites and merozoites of *E. tenella* (Fig 3C). Thus the parasite likely presents a complex set of antigens to the host, much like *T. gondii* SRSs, rather than a single one like *Plasmodium falciparum* PfEMP1. Analysis of stage-specific *sag* expression in populations of clonal *E. tenella* suggests that the expressed repertoire is most complex in second- generation merozoites (Fig 3D). A small number of *sagA* genes peak in expression at each stage, *sagB* genes all peak in expression in second generation merozoites suggesting that they may be particularly important during the later, pathogenic, stages of infection (Fig 3D).

The total number of *sag* genes varies greatly between species (Table 1). This may simply reflect the overall phylogeny (see Fig 1A) but it is notable that species that cause the most severe pathologies have higher numbers of *sags*. Thus *E. praecox*, which causes only superficial damage and is widely regarded as the least pathogenic (Allen and Jenkins 2010) has only 19 *sags*, whilst *E. tenella, E. necatrix* and *E. brunetti*, which develop deep in the mucosa causing tissue damage, inflammation and intestinal haemorrhage (McDonald and Shirley 1987), have 89, 119 and 105 *sags* respectively. However *E. mitis* with the greatest number of *sags* (172) does not fit this pattern; like all species it can impair bird performance and productivity but does not cause gross lesions in the intestine. It is also the case that species which induce potent immunity against re-infection after exposure to small numbers of parasites (*E. maxima*, *E. praecox* and *E. acervulina*) have the lowest numbers of *sags* whilst those that are least immunogenic (*E. necatrix* and *E. tenella*) have high numbers; however *E. brunetti* and *E. mitis*, which are of intermediate immunogenicity, do not fit this pattern.

We used remote homology detection to explore the evolutionary origin of the *sag* family. The cysteine-rich secretory protein family (CAP), found in a wide range of eukaryotes (Gibbs et al. 2008), had low but significant sequence similarity to *sags* (Fig 3E). The most similar CAP-domain containing proteins were those of *T. gondii* suggesting that *sag* genes are likely derived from CAP-domain containing genes in the common ancestor of *E. tenella* and *T. gondi,* rather than by horizontal transfer from another species such as the host. Recent protein structural evidence shows that the *srs* surface antigen gene family in *T. gondii* is related to a small group of cysteine-rich proteins in *Plasmodium* (Arredondo et al. 2012). These results suggest distinct evolutionary origins for the principle surface antigen gene families of the relatively closely related *T. gondii* and *E. tenella*.

Discussion

Control of pathogens such as *Eimeria* species has been essential for development of modern poultry production and is increasingly important for providing global food security. The availability of genomic resources for the seven *Eimeria*

species that infect domestic chickens will underpin development and longevity of new anticoccidial drugs and vaccines.

The most striking feature of the *Eimeria* genomes is the disruption of more than half of all protein coding sequences with HAARs. Across every chromosome of each species we found regions where 20-30% of the sequence comprised simple tandem repeats, interspersed with relatively repeat-free regions. Although a variety of different simple repeats were found outside of coding regions, those within coding regions were almost always based on runs of the trinucleotide CAG. We examined whether these repeats might have a particular function in the proteome but could find no association with known functional groups and showed that the repeats localized to structurally neutral regions within proteins. We hypothesise that poly-CAG is the most benign and easily evolvable coding repeat and that there is sufficient pressure on maintaining repeat banding to allow for frequent deleterious mutations. The repeat regions might function at the DNA level, perhaps in tertiary structure or gene regulation, and be selected for functional neutrality at the protein level, as we have observed. This hypothesis could be explored by chromatin immunoprecipitation and chromosome conformation capture studies. An alternative hypothesis is that rapid mutation of coding sequences provides evolutionary advantages, which would suggest selection for repeats at the protein level. Although the HAARs we examined appeared to be neutral in their effects on protein structures (presumably deleterious changes are lost in the population) occasional mutations may have been sufficiently beneficial to support the heavy burden. However genes known to be involved in host-parasite interaction were found to be relatively protected from HAARs, and why repeats should have appeared in a banded pattern across chromosomes is not clear in this scenario.

We have characterised, for the first time, retrotransposon-like elements in an apicomplexan. Disruption or replacement of genes by transposon-mediated transgenesis holds the potential to improve our molecular arsenal for unraveling parasite biology, and also for developing attenuated vaccines. This approach has been attempted in *Eimeria* species using piggyBAC transposons, however, rates

of insertion of these elements are low and target particular sequences (Su et al. 2012). Random, high frequency insertions of native transposons would allow high throughput knockout screens, accelerating our understanding of *Eimeria* biology. The retrotransposons identified here are at best partially degraded and it is not clear whether they are actively retrotransposed. However, it opens the possibility that intact retrotransposons might be present in more distantly related *Eimeria* species.

To develop cheap, effective new vaccines for coccidiosis we must understand how the parasites interact with the host immune system. Key to this may be the *sag* family of genes encoding surface antigens, some of which have been shown *in vitro* to induce pro-inflammatory cytokine responses (Chow et al. 2011). Whilst some *sags* are shared across all the *Eimeria* species studied, others are clade-specific. We found multiple *sags* to be co-expressed on the surface of infective parasites suggesting that the avian immune system is presented with a diverse array of related epitopes, which could potentially aggravate inflammation. Parasites that rely on a healthy host for vector transmission are thought to use diverse arrays of related antigens to reduce pathogenicity (Spence et al. 2013) but for *Eimeria*, which has a simple oral-fecal life cycle, the induction of inflammation leading to diarrhea could increase parasite transmission. *Eimeria sag* genes have evolved from the CAP domain superfamily of cysteine-rich secretory proteins and are unrelated to the major surface antigens of other Apicomplexa. CAP domains are also found on the surface of parasitic helminths where they are proposed to interact with host immune systems (Chalmers and Hoffmann 2012).

Methods

Parasite cultivation

The Houghton (H) strains of *E. acervulina*, *E. brunetti*, *E. mitis*, *E. necatrix*, *E. praecox* and *E. tenella*, and the Weybridge (W) strain of *E. maxima*, were used for principle sequencing. The H strains were isolated at the Houghton Poultry Research Station (UK) and are the progeny of single oocyst infections. The *E. maxima* W strain was isolated at the Weybridge Central Veterinary Laboratory from a single oocyst infection. The Wisconsin (Wis) (McDougald and Jeffers 1976) and Nippon-2 (Nt2) *E. tenella* lines were used for comparative analyses. All parasites were propagated *in vivo* in three to seven week old Light Sussex chickens under specific pathogen free (SPF) conditions at the Institute for Animal Health or Royal Veterinary College and purified using established methods (Long et al. 1976).

Summary of genome sequencing and assembly

Genomic DNA was prepared from purified sporulated oocysts. We used Sanger capillary sequencing data previously generated for *E. tenella*, described in Ling et al (2007) and deposited in the Trace Archive (http://www.ncbi.nlm.nih.gov/Traces; CENTER_PROJECT = "EIMER"). These paired reads had a range of insert sizes to a combined coverage of ~8x. *E. tenella* Illumina GAIIx sequencing libraries were prepared with insert sizes of 300 bp and 3 kbp and either 54 bp or 76 bp paired-end reads with a combined coverage of ~160x. Capillary reads were assembled using ARACHNE v3.2 using default parameters (Batzoglou et al. 2002). IMAGE (Tsai et al. 2010) was used to fill gaps in scaffolds and extend contigs with Illumina reads, running 6 iterations (3 with k-mer=31 and 3 with k-mer=27), mapping with BWA (Li and Durbin 2009). The consensus sequence from the ARACHNE-IMAGE assembly was corrected with Illumina reads using iCORN (Otto et al. 2010). All Illumina reads which did not map to this assembly were assembled using Velvet (Zerbino and Birney 2008) and these contigs were added to the final assembly.

For tier 2 (*E. necatrix* H, *E. maxima* W and *E. acervulina* H) and tier 3 (*E. mitis* H, *E. praecox* H and *E. brunetti* H) genomes 500bp Illumina TruSeq libraries were

prepared and sequenced as 76bp paired-end reads on an Illumina HiSeq 2000 platform to a depth of 199x theoretical genome coverage for *E. acervulina*, 288x for *E. maxima*, 559x for *E. necatrix*, 143x for *E. brunetti*, 520x for *E. mitis* and 102x for *E. praecox* as in (Kozarewa et al. 2009). Reads were assembled using Velvet (Zerbino and Birney 2008), scaffolding was performed with SSPACE (Boetzer et al. 2011) and gaps were filled using IMAGE (Tsai et al. 2010).

Full methodological details of genome sequencing, assembly and annotation are provided in the supplement.

Whole genome phylogeny

Each of 814 one-to-one orthologue groups shared across the seven *Eimeria* species with *Toxoplasma gondii* ME49 were aligned using MAFFT v7 (Katoh and Standley 2013). Highly variable sites were trimmed using trimal (Capella-Gutierrez et al. 2009) ("automated1" option). The alignments were concatenated using FASconCAT (Kuck and Meusemann 2010) and the resulting alignment used to construct a maximum likelihood phylogenetic tree using RAxML with the model PROTGAMMALG4XF (Stamatakis et al. 2005; Le et al. 2012), bootstrap n=100. Bootstrap percentage support is shown along the branching nodes. The tree was rooted using *T. gondii* as the outgroup.

Transcriptome sequencing and analysis of *Eimeria tenella*

Unsporulated oocysts (two biological replicates), sporulated oocysts (single replicate), purified sporozoites (single replicate) and second generation merozoites (single replicate) of the *E. tenella* H strain were selected for transcriptome analysis after harvest and purification as described previously (Novaes et al. 2012). Total RNA was extracted from oocysts using a Qiagen RNeasy kit (Qiagen, Crawley, UK) and from sporozoites and second-generation merozoites using the Qiagen RNeasy Animal Cells purification protocol as described by the manufacturer. All samples were DNase treated using the Qiagen RNase-free DNase kit during RNA purification.

Library preparation, sequencing protocols and data processing were the same as for (Reid et al. 2012). For *sag* gene expression clustering in Fig. 3D, mean RPKM values were taken for unsporulated oocyst samples and clustering performed using MBCluster.seq with 10 clusters (Si et al. 2014). Clusters were ordered by the stage of peak expression across each cluster. The figure was drawn with Circos (Krzywinski et al. 2009).

Synteny analysis

MCScanX (Wang et al. 2012) was used to determine regions of synteny between pairs of species based on the order of pairwise orthologues identified using OrthoMCL (Li et al. 2003). Default values were used except for the MATCH_SIZE option, which was set to 3 for comparison of *E. tenella* and *T. gondii*. For other comparisons it was set to 5.

Classification and analysis of gene families

Methodological details are available in the supplement.

Retrotransposon analysis

Methodological details are available in the supplement.

Single-cell expression of *sag* genes

Single fluorescent *E. tenella* sporozoites and second-generation merozoites (Clark et al. 2008) were sorted into 96-well plates. RNA was reverse transcribed for a panel of target transcripts by adding gene-specific reverse primers (Supplemental Table S9). Stage-specific multi-cell derived cDNA or single cell cDNA with each single-plex primer pair were included as positive controls P1 and P2. No template multiplex and no reverse transcription reactions were included for each assay as negative controls N1 and N2.

MDBK cell binding by SAGs

Confluent monolayers of MDBK cells were blocked with 1% BSA in PBS for 2 h at 4^{o}C, washed three times in PBS then incubated with recombinant-expressed EtSAG proteins (0.5 mg/ml) for 1h at 40^{o}C. Monolayers were washed four times

with PBS to remove unbound proteins, then cells and bound proteins were solublised in SDS sample buffer, separated by SDS-PAGE, transferred to nitrocellulose by electroblotting and probed with rabbit hyperimmune sera raised to recombinant SAG proteins.

Classification of repeat regions

We used Tandem Repeat Finder (Benson 1999) to identify short tandem repeats in genomic sequences. It was run with the following parameters: 2, 1000, 1000, 80, 10, 25, 1000, as were used in (Ling et al. 2007) and processed using TRAP v1.1 (Sobreira et al. 2006), run with default parameters. We defined a repeat density of >= 0.05 over a window of 1kb as defining a repeat-rich region based on the distribution of repeat density across *E. tenella*. Scaffolds shorter than 5kb were not included in the analysis.

We used SEG (Wootton and Federhen 1996) with parameters 7, 0, 0, –l, to identify Homopolymeric Amino Acid Repeats (HAARs) of length 7 or greater. Repeats of 'X', the symbol used when the amino acid is unknown, and repeats which contained stop codons were removed from the SEG output file. To determine the HAAR content of protein sequences in other organisms we ran SEG on all complete proteomes from UniProt (~2000 species, ~10m sequences). We then ranked each species by the median number of repeats per sequence. We took the top ten, excluding viruses, and also *T. gondii* for comparison.

Full methodological details of protein structural and proteomic analyses of HAARs are available in the supplement.

Data Access

Sequencing reads and assemblies for each genome are available from the European Nucleotide Archive (http://www.ebi.ac.uk/ena/): *E. tenella* Houghton, *E. maxima* Weybridge, *E. acervulina* Houghton, *E necatrix* Houghton, *E. mitis* Houghton, *E. brunetti* Houghton and *E. praecox* Houghton (PRJEB4918). Sequencing reads for *E. tenella* Nippon and *E. tenella* Wisconsin are also available from the ENA (PRJEB4009). Transcriptome sequences for *E. tenella* Houghton

are archived in the ENA (ERP001847) and in ArrayExpress (https://www.ebi.ac.uk/arrayexpress/; E-ERAD-109).

Acknowledgements

The work was funded by BBSRC grants S17413 and S19705/6, Wellcome Trust core funding to Wellcome Trust Sanger Institute (WTSI) and faculty baseline funding from the King Abdullah University of Science and Technology (KAUST). We acknowledge the support of Mike Quail, Karen Mungall, Carol Churcher and members of the core DNA pipelines for sequencing at WTSI and Bioscience core laboratories for sequencing operations at KAUST. We acknowledge Martin Aslett from WTSI for data submission and Dora Harvey and Fionnadh Carroll from the Institute for Animal Health for technical assistance, Kiew-Lan Wan would like to acknowledge funding from the Ministry of Science, Technology and Innovation, Malaysia (Project No. 07-05-16-MGI-GMB10) and the Universiti Kebangsaan Malaysia (Project No. DIP-2012-21). JP and SSH were funded by the Canadian Institute for Health Research (CIHR #MOP84556).

Disclosure

The authors declare no conflicts of interest.

Figure legends

Figure 1. Whole genome phylogeny and synteny between Eimeria species. (A) Maximum likelihood phylogeny showing the evolutionary relationships between *Eimeria* species based on alignment of 814 one-to-one orthologs shared with *T. gondii*. The scale is in substitutions per site. (B) Genomic scaffolds were placed onto optical maps. Black bands show map coverage. Coverage was noticeably better for *E. tenella* than the tier 2 species. *E. tenella* maps are named as chromosomes (e.g. C1) where it was possible to reliably identify that chromosome, otherwise they are given their optical map numbers (e.g. 04). Each *E. tenella* map has been assigned a colour and ribbons highlight syntenic regions in the related genomes. *E. necatrix* is most closely related to *E. tenella* and correspondingly shows the greatest degree of synteny. The clearest exceptions are (1) 010, which is split between two optical contigs in *E. necatrix* and (2) 08, which is similarly split. Map coverage is lower in *E. acervulina* and *E. maxima* and this gives the impression that there is a great deal of novel sequence in these species. However, this is largely the result of differential representation of the genomes in their respective maps. Each map is annotated with repeat poor (blue) and repeat rich (red) regions ≥30kb. This highlights the barcode-like patterning across the whole of each genome.

Figure 2. Characterisation of Homopolymeric Amino Acid Repeats (HAARs) in *Eimeria* protein sequences. (A) *Eimeria tenella* has a greater number of HAARs than any other genome sequenced and a distinct distribution of HAAR types compared to other repeat-rich genomes including *Plasmodium falciparum* and the more closely related and not especially repeat-rich *Toxoplasma*.
(B) The most common short tandem repeats in *Eimeria* genomes are variations on CAG. The second most common are variations on a telomere repeat which we call telomere-like repeats due to their locations throughout the genome. (C) CAG repeats occur in protein-coding regions of the genome more than expected. (D) CAG repeats can encode strings of one of five amino acids. In *Eimeria* they tend to encode alanine and glutamine more often than expected, serine as often as expected and leucine and cysteine more rarely than expected. A very similar

pattern is observed in a limited selection of *E. tenella* peptides derived from proteomics experiments.

Figure 3. Analysis of the principle family of surface antigens in *Eimeria spp*. (A) The four loci of tandemly repeated *sag* genes in *E. tenella* are shown with each *sag* gene represented by bars describing the relative expression levels in four stages of the lifecycle. Gene names used previously in the literature are shown where appropriate. Arrows indicate direction of transcription. (B) We found that, of those tested, SAG proteins from subfamily A bound host cells, but those from family B did not. U= Unwashed, W = Washed, B = Bound. (C) Expression of multiple sag genes in individual cells of *E. tenella* was detected using RT-PCR. Ten cells were analysed for each of eight genes in both sporozoites and second-generation merozoites. 1-10 = single cell multiplex test RT-PCRs. Controls: P1 = positive, cDNA library multiplex RT-PCR, N1 = negative, no template multiplex RT-PCR, P2 = positive, single sporozoite, single target RT-PCR, N2 = negative, single sporozoite, single target PCR with no RT. (D) Expression values for all *E. tenella sag* genes were clustered and those clusters ordered by mean peak expression, showing that most genes peak in the second generation merozoite. Where appropriate genes are annotated with their genomic locus as defined in A. (E) Patterns of homology for *Eimeria sag* and *Toxoplasma srs* genes suggest that while *Toxoplasma* acquired the precursors to its key family of 6-cys surface antigens from a horizontal gene transfer of metazoan ephrin, *Eimeria* has derived them from the Cysteine-rich secretory proteins, Antigen 5, and Pathogenesis-related 1 protein (CAP) family already found in Apicomplexa. Thin, single-headed arrows show phylogenetic paths for CAP and ephrin-related domains, while bold arrows show the best evidence for the closest relatives of the *sag* and *srs* families. Numbers indicate gene frequencies for each family.

Tables

Table 1. *Eimeria* species biology and genomic *sag* repertoire. Numbers of *sag* genes in each *Eimeria* genome, their breakdown into subfamilies and the numbers of *sag* pseudogenes (*E. tenella*) or pseudogene fragments (other species).
SI = small intestine, M = malabsorptive disease, H = haemorrhagic disease. Levels of pathogenicity are derived from (Long et al. 1976).
*Gametogony occurs in the caeca.

Species	Site of development	Disease type	Pathogenicity	sagA	sagB	sagC	Total	Pseudogene fragments
E. praecox H	SI (upper)	M	+	15	0	4	19	20
E. maxima W	SI (mid)	M	+++	35	0	4	39	29
E. acervulina H	SI (upper)	M	++	13	1	2	16	16
E. brunetti H	SI (lower), rectum, caeca	H	++++	61	0	44	105	39
E. mitis H	SI (lower)	M	++	145	0	27	172	128
E. necatrix H	SI (mid), caeca*	H	+++++	86	32	1	119	102
E. tenella H	Caeca	H	++++	60	28	1	89	23

References

Allen PC, Jenkins MC. 2010. Observations on the gross pathology of Eimeria praecox infections in chickens. *Avian diseases* **54**(2): 834-840.

Anamika, Srinivasan N, Krupa A. 2005. A genomic perspective of protein kinases in Plasmodium falciparum. *Proteins* **58**(1): 180-189.

Arredondo SA, Cai M, Takayama Y, MacDonald NJ, Anderson DE, Aravind L, Clore GM, Miller LH. 2012. Structure of the Plasmodium 6-cysteine s48/45 domain. *Proc Natl Acad Sci U S A* **109**(17): 6692-6697.

Balaji S, Babu MM, Iyer LM, Aravind L. 2005. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res* **33**(13): 3994-4006.

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES. 2002. ARACHNE: a whole-genome shotgun assembler. *Genome Res* **12**(1): 177-189.

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

Blake DP, Billington KJ, Copestake SL, Oakes RD, Quail MA, Wan KL, Shirley MW, Smith AL. 2011. Genetic mapping identifies novel highly protective antigens for an apicomplexan parasite. *PLoS Pathog* **7**(2): e1001279.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**(4): 578-579.

Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M. 2010. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog* **6**(10): e1001165.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**(15): 1972-1973.

Chalmers IW, Hoffmann KF. 2012. Platyhelminth Venom Allergen-Like (VAL) proteins: revealing structural diversity, class-specific features and biological associations across the phylum. *Parasitology* **139**(10): 1231-1245.

Chapman HD, Barta JR, Blake D, Gruber A, Jenkins M, Smith NC, Suo X, Tomley FM. 2013. A selective review of advances in coccidiosis research. *Advances in parasitology* **83**: 93-171.

Chow YP, Wan KL, Blake DP, Tomley F, Nathan S. 2011. Immunogenic Eimeria tenella glycosylphosphatidylinositol-anchored surface antigens (SAGs) induce inflammatory responses in avian macrophages. *PLoS One* **6**(9): e25233.

Clark JD, Billington K, Bumstead JM, Oakes RD, Soon PE, Sopp P, Tomley FM, Blake DP. 2008. A toolbox facilitating stable transfection of Eimeria species. *Mol Biochem Parasitol* **162**(1): 77-86.

Coulson RM, Hall N, Ouzounis CA. 2004. Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum. *Genome Res* **14**(8): 1548-1554.

del Cacho E, Pages M, Gallego M, Monteagudo L, Sanchez-Acedo C. 2005. Synaptonemal complex karyotype of Eimeria tenella. *Int J Parasitol* **35**(13): 1445-1451.

Fentress SJ, Behnke MS, Dunay IR, Mashayekhi M, Rommereim LM, Fox BA, Bzik DJ, Taylor GA, Turk BE, Lichti CF et al. 2010. Phosphorylation of immunity-related GTPases by a Toxoplasma gondii-secreted kinase promotes macrophage survival and virulence. *Cell Host Microbe* **8**(6): 484-495.

Gibbs GM, Roelants K, O'Bryan MK. 2008. The CAP superfamily: cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins--roles in reproduction, cancer, and immune defense. *Endocrine reviews* **29**(7): 865-897.

He CY, Shaw MK, Pletcher CH, Striepen B, Tilney LG, Roos DS. 2001. A plastid segregation defect in the protozoan parasite Toxoplasma gondii. *EMBO J* **20**(3): 330-339.

Jahn D, Matros A, Bakulina AY, Tiedemann J, Schubert U, Giersberg M, Haehnel S, Zoufal K, Mock HP, Kipriyanov SM. 2009. Model structure of the immunodominant surface antigen of Eimeria tenella identified as a target for sporozoite-neutralizing monoclonal antibody. *Parasitology research* **105**(3): 655-668.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**(4): 772-780.

Kordis D. 2005. A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. *Gene* **347**(2): 161-173.

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**(4): 291-295.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**(9): 1639-1645.

Kuck P, Meusemann K. 2010. FASconCAT: Convenient handling of data matrices. *Molecular phylogenetics and evolution* **56**(3): 1115-1118.

Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol* **29**(10): 2921-2936.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**(9): 2178-2189.

Liberator P, Anderson J, Feiglin M, Sardana M, Griffin P, Schmatz D, Myers RW. 1998. Molecular cloning and functional expression of mannitol-1-phosphatase from the apicomplexan parasite Eimeria tenella. *J Biol Chem* **273**(7): 4237-4244.

Lim L, McFadden GI. 2010. The evolution, metabolism and functions of the apicoplast. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **365**(1541): 749-763.

Ling KH, Rajandream MA, Rivailler P, Ivens A, Yap SJ, Madeira AM, Mungall K, Billington K, Yee WY, Bankier AT et al. 2007. Sequencing and analysis of chromosome 1 of Eimeria tenella reveals a unique segmental organization. *Genome Res* **17**(3): 311-319.

Long PL, Millard BJ, Joyner LP, Norton CC. 1976. A guide to laboratory techniques used in the study and diagnosis of avian coccidiosis. *Folia veterinaria Latina* **6**(3): 201-217.

McDonald V, Shirley MW. 1987. The endogenous development of virulent strains and attenuated precocious lines of Eimeria tenella and E. necatrix. *J Parasitol* **73**(5): 993-997.

McDougald LR, Jeffers TK. 1976. Eimeria tenella (Sporozoa, Coccidia): Gametogony following a single asexual generation. *Science* **192**(4236): 258-259.

Miranda-Saavedra D, Gabaldon T, Barton GJ, Langsley G, Doerig C. 2012. The kinomes of apicomplexan parasites. *Microbes Infect* **14**(10): 796-810.

Novaes J, Rangel LT, Ferro M, Abe RY, Manha AP, de Mello JC, Varuzza L, Durham AM, Madeira AM, Gruber A. 2012. A comparative transcriptome analysis reveals expression profiles conserved across three Eimeria spp. of domestic fowl and associated with multiple developmental stages. *Int J Parasitol* **42**(1): 39-48.

Ogedengbe JD, Hanner RH, Barta JR. 2011. DNA barcoding identifies Eimeria species and contributes to the phylogenetics of coccidian parasites (Eimeriorina, Apicomplexa, Alveolata). *Int J Parasitol* **41**(8): 843-850.

Otto TD, Sanders M, Berriman M, Newbold C. 2010. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26**(14): 1704-1707.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**(9): 1061-1067.

Peixoto L, Chen F, Harb OS, Davis PH, Beiting DP, Brownback CS, Ouloguem D, Roos DS. 2010. Integrative genomic approaches highlight a family of parasite-specific kinases that regulate host responses. *Cell Host Microbe* **8**(2): 208-218.

Perutz MF, Pope BJ, Owen D, Wanker EE, Scherzinger E. 2002. Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid beta-peptide of amyloid plaques. *Proc Natl Acad Sci U S A* **99**(8): 5596-5600.

Reid AJ, Vermont SJ, Cotton JA, Harris D, Hill-Cawthorne GA, Konen-Waisman S, Latham SM, Mourier T, Norton R, Quail MA et al. 2012. Comparative genomics of the apicomplexan parasites Toxoplasma gondii and Neospora caninum: Coccidia differing in host range and transmission strategy. *PLoS Pathog* **8**(3): e1002567.

Saeij JP, Coller S, Boyle JP, Jerome ME, White MW, Boothroyd JC. 2007. Toxoplasma co-opts host gene expression by injection of a polymorphic kinase homologue. *Nature* **445**(7125): 324-327.

Schmatz DM, Baginsky WF, Turner MJ. 1989. Evidence for and characterization of a mannitol cycle in Eimeria tenella. *Mol Biochem Parasitol* **32**(2-3): 263-270.

Shirley MW, Smith AL, Tomley FM. 2005. The biology of avian Eimeria with an emphasis on their control by vaccination. *Advances in parasitology* **60**: 285-330.

Si Y, Liu P, Li P, Brutnell TP. 2014. Model-based clustering for RNA-seq data. *Bioinformatics* **30**(2): 197-205.

Sobreira TJ, Durham AM, Gruber A. 2006. TRAP: automated classification, quantification and annotation of tandemly repeated sequences. *Bioinformatics* **22**(3): 361-362.

Spence PJ, Jarra W, Levy P, Reid AJ, Chappell L, Brugat T, Sanders M, Berriman M, Langhorne J. 2013. Vector transmission regulates immune control of Plasmodium virulence. *Nature* **498**(7453): 228-231.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**(4): 456-463.

Su H, Liu X, Yan W, Shi T, Zhao X, Blake DP, Tomley FM, Suo X. 2012. piggyBac transposon-mediated transgenesis in the apicomplexan parasite Eimeria tenella. *PLoS One* **7**(6): e40075.

Tabares E, Ferguson D, Clark J, Soon PE, Wan KL, Tomley F. 2004. Eimeria tenella sporozoites and merozoites differentially express glycosylphosphatidylinositol-anchored variant surface proteins. *Mol Biochem Parasitol* **135**(1): 123-132.

Talevich E, Kannan N. 2013. Structural and evolutionary adaptation of rhoptry kinases and pseudokinases, a family of coccidian virulence factors. *BMC evolutionary biology* **13**: 117.

Talevich E, Mirza A, Kannan N. 2011. Structural and evolutionary divergence of eukaryotic protein kinases in Apicomplexa. *BMC evolutionary biology* **11**: 321.

Tsai IJ, Otto TD, Berriman M. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* **11**(4): R41.

van Dooren GG, Tomova C, Agrawal S, Humbel BM, Striepen B. 2008. Toxoplasma gondii Tic20 is essential for apicoplast protein import. *Proc Natl Acad Sci U S A* **105**(36): 13574-13579.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**(7): e49.

Ward P, Equinet L, Packer J, Doerig C. 2004. Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote. *BMC Genomics* **5**: 79.

Williams RB. 1998. Epidemiological aspects of the use of live anticoccidial vaccines for chickens. *Int J Parasitol* **28**(7): 1089-1098.

Wootton JC, Federhen S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods in enzymology* **266**: 554-571.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**(5): 821-829.

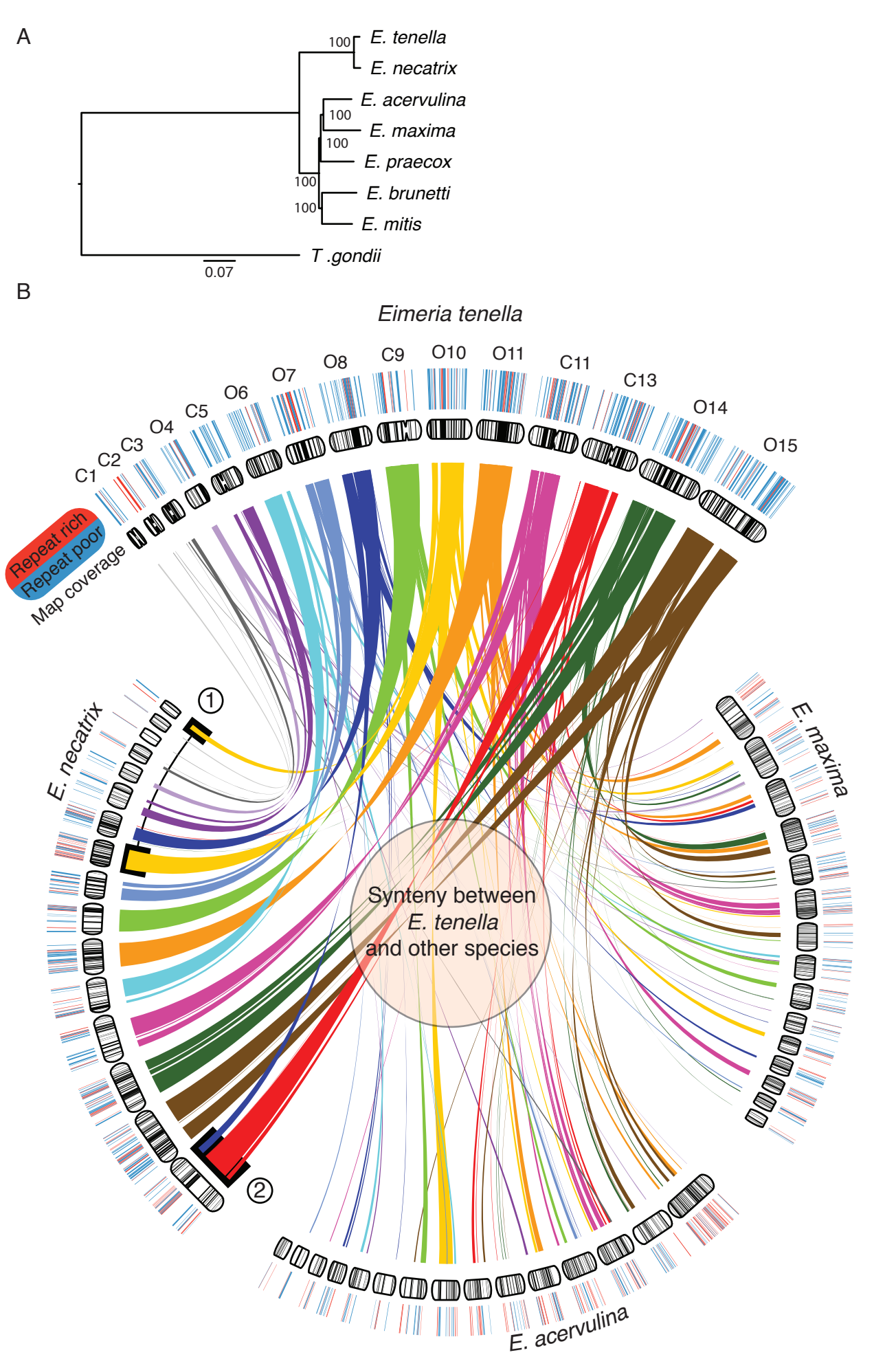

A
100
E. tenella
E. necatrix
E. acervulina
100
E. maxima
100
E. praecox
100
E. brunetti
100
E. mitis
T .gondii
0.07
B
Eimeria tenella
C1
C2
C3
O4
C5
O6
O7
O8
C9
O10
O11
C11
C13
O14
O15
Repeat rich
Repeat poor
Map coverage
E. necatrix
E. maxima
E. acervulina
Synteny between
E. tenella
and other species
1
2

Figure 1

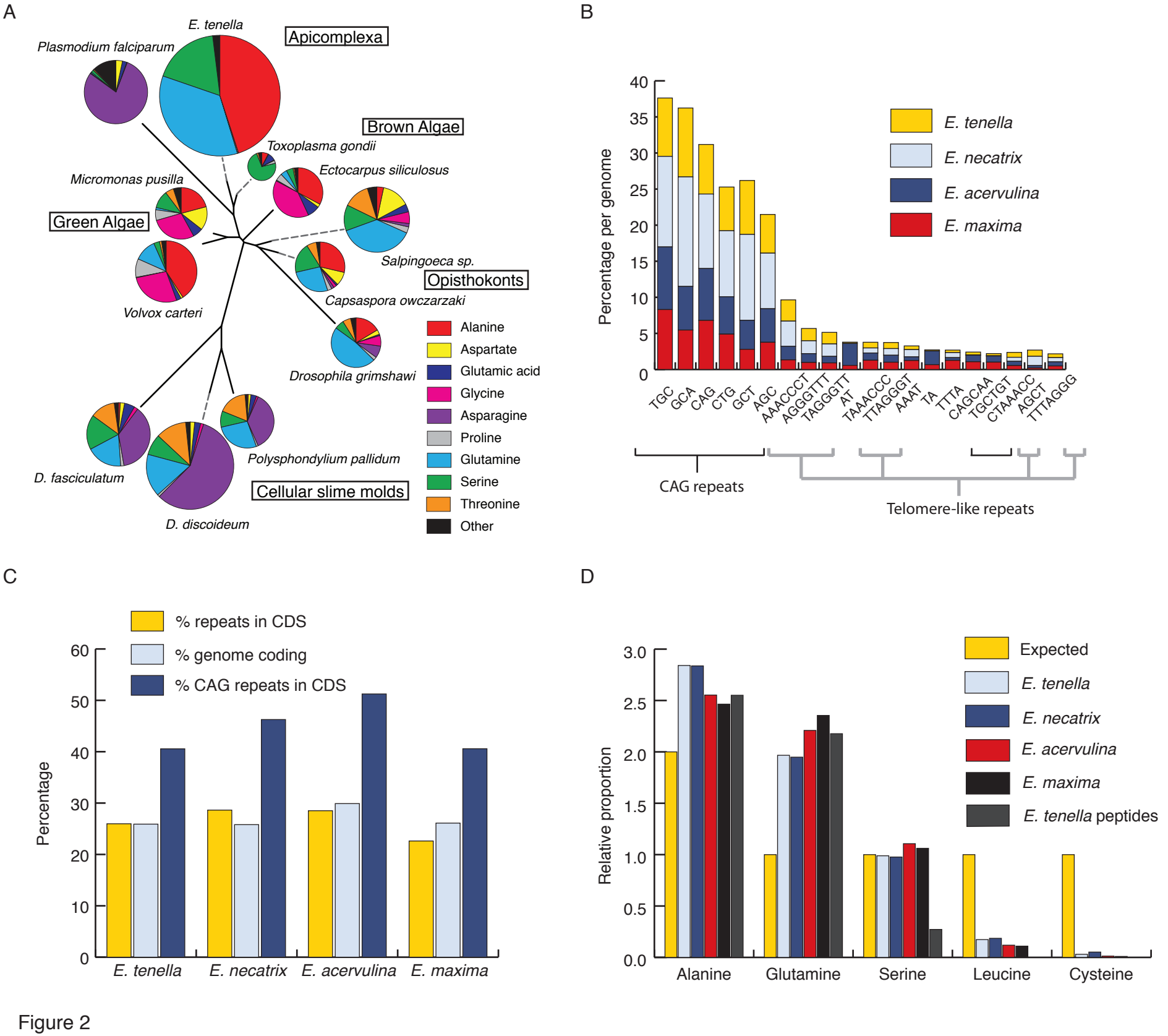
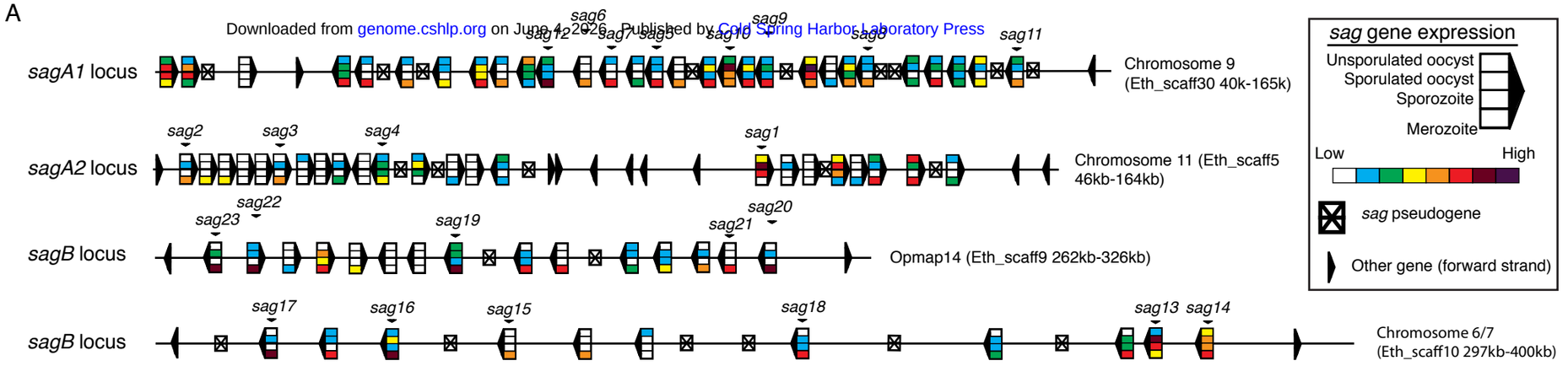
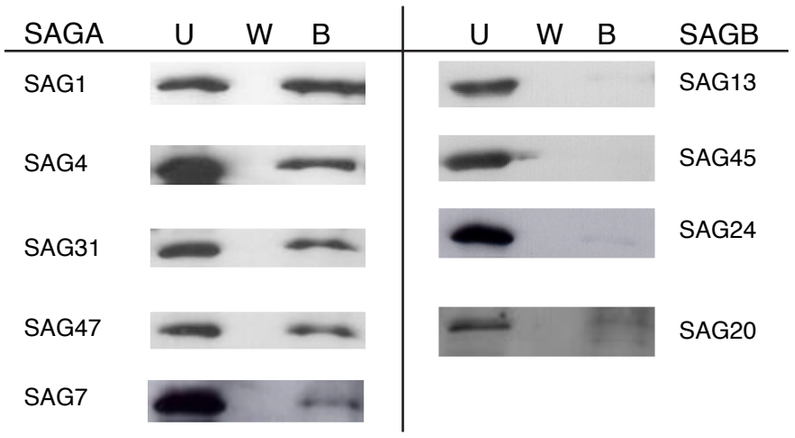
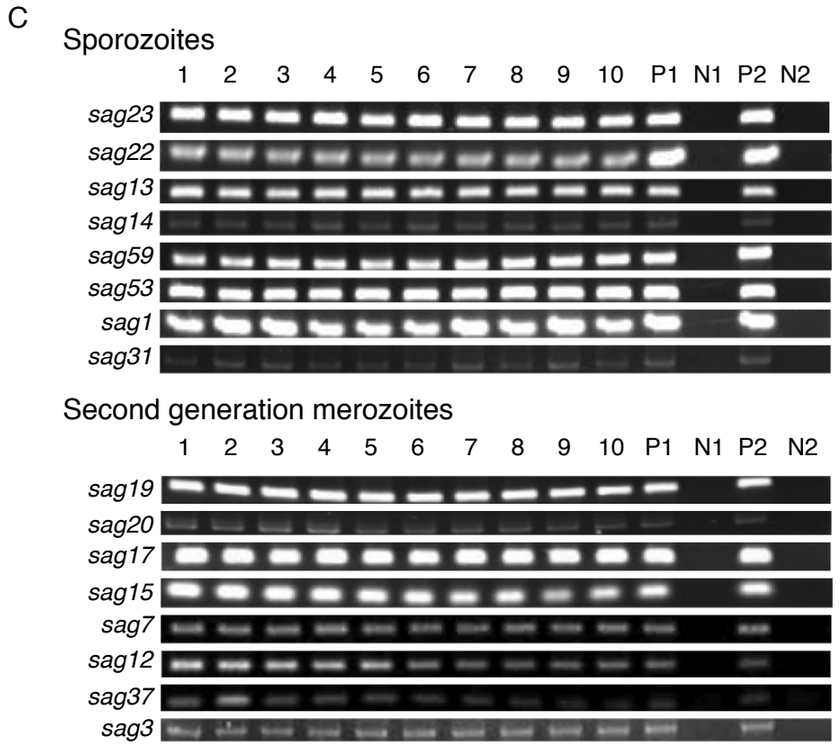
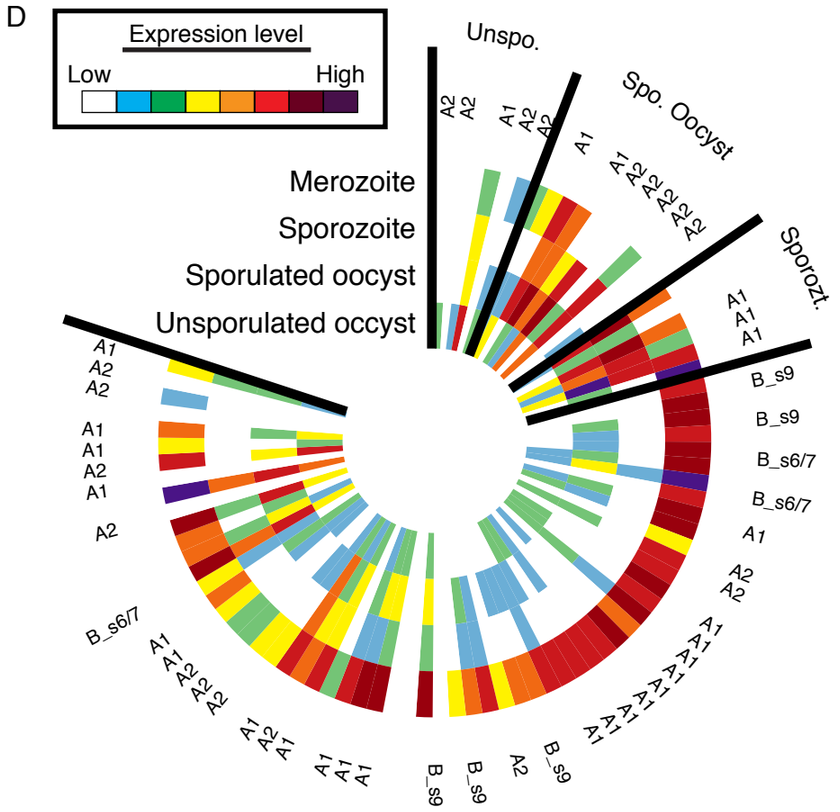

A
E. tenella
Plasmodium falciparum
Apicomplexa
Brown Algae
Toxoplasma gondii
Ectocarpus siliculosus
Micromonas pusilla
Green Algae
Salpingoeca sp.
Opisthokonts
Capsaspora owczarzaki
Volvox carteri
Drosophila grimshawi
Alanine
Aspartate
Glutamic acid
Glycine
Asparagine
Proline
Glutamine
Serine
Threonine
Other
Polysphondylium pallidum
D. fasciculatum
Cellular slime molds
D. discoideum
B
Percentage per genome
E. tenella
E. necatrix
E. acervulina
E. maxima
TGC
GCA
CAG
CTG
GCT
AGC
AAACCCT
AGGGTTT
TAGGGTT
AT
TAAACCC
TTAGGGT
AAAT
TA
TTTA
CAGCAA
TGCTGT
CTAAACC
AGCT
TTTAGGG
CAG repeats
Telomere-like repeats
C
% repeats in CDS
% genome coding
% CAG repeats in CDS
Percentage
E. tenella
E. necatrix
E. acervulina
E. maxima
D
Relative proportion
Expected
E. tenella
E. necatrix
E. acervulina
E. maxima
E. tenella peptides
Alanine
Glutamine
Serine
Leucine
Cysteine

Figure 2









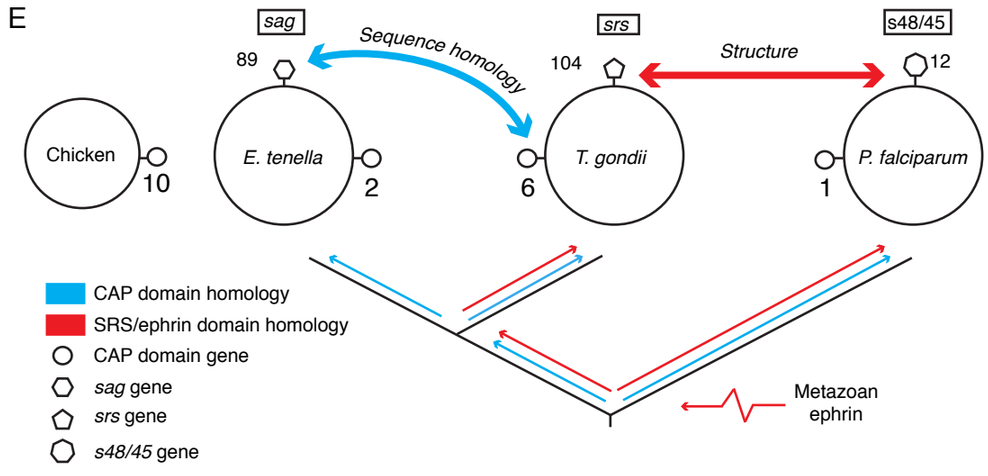


Figure 3