MICROBIAL GENOMICS

RESEARCH ARTICLE

Li *et al.*, *Microbial Genomics* 2021;7:000711

DOI 10.1099/mgen.0.000711

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

# *Cryptosporidium felis* differs from other *Cryptosporidium* spp. in codon usage

Jiayu Li[1,2]†, Yaqiong Guo[1]†, Dawn M. Roellig[3], Na Li[1], Yaoyu Feng[1,2,*] and Lihua Xiao[1,2,*]

**Abstract**

*Cryptosporidium* spp. are important enteric pathogens in a wide range of vertebrates including humans. Previous comparative analysis revealed conservation in genome composition, gene content, and gene organization among *Cryptosporidium* spp., with a progressive reductive evolution in metabolic pathways and invasion-related proteins. In this study, we sequenced the genome of zoonotic pathogen *Cryptosporidium felis* and conducted a comparative genomic analysis. While most intestinal *Cryptosporidium* species have similar genomic characteristics and almost complete genome synteny, fewer protein-coding genes and some sequence inversions and translocations were found in the *C. felis* genome. The *C. felis* genome exhibits much higher GC content (39.6%) than other *Cryptosporidium* species (24.3–32.9%), especially at the third codon position (GC3) of protein-coding genes. Thus, *C. felis* has a different codon usage, which increases the use of less energy costly amino acids (Gly and Ala) encoded by GC-rich codons. While the tRNA usage is conserved among *Cryptosporidium* species, consistent with its higher GC content, *C. felis* uses a unique tRNA for GTG for valine instead of GTA in other *Cryptosporidium* species. Both mutational pressures and natural selection are associated with the evolution of the codon usage in *Cryptosporidium* spp., while natural selection seems to drive the codon usage in *C. felis*. Other unique features of the *C. felis* genome include the loss of the entire traditional and alternative electron transport systems and several invasion-related proteins. Thus, the preference for the use of some less energy costly amino acids in *C. felis* may lead to a more harmonious parasite–host interaction, and the strengthened host-adaptation is reflected by the further reductive evolution of metabolism and host invasion-related proteins.

## DATA SUMMARY

The whole-genome sequence data of *Cryptosporidium felis* have been deposited in the National Centre for Biotechnology Information (NCBI) under BioProject accession PRJNA640950, including Sequence Read Archive (SRA) under accessions SRR12064568 and SRR12064569, and genome assembly with full annotations under GenBank accession JABXOJ000000000.

## INTRODUCTION

*Cryptosporidium* spp. are apicomplexan parasites that inhabit the gastrointestinal tract of humans and various animals, causing moderate-to-severe diarrhoea. To date, there are nearly 40 known *Cryptosporidium* species and about the same number of genotypes of uncertain species status [1]. Among them, *C. parvum* and *C. hominis* are the most common ones responsible for human cryptosporidiosis, followed by *C. meleagridis*, *C. felis*, and *C. canis* [2]. Some other *Cryptosporidium* species are found at lower frequency in humans [1]. Between the two most common human-pathogenic species, *C. hominis* mainly infects humans, nonhuman primates, and equine animals, while *C. parvum* can infect these animals as well as ruminants and rodents [1]. In contrast, *C. felis* is a host-adapted species, with cats being the only major known host beyond humans.

The genomes of several *Cryptosporidium* species have been sequenced, including intestinal species *C. parvum*, *C. hominis*, *C. cuniculus*, *C. tyzzeri*, *C. meleagridis*, *C. viatorum*, *C. ubiquitum*, *C. bovis*, *C. ryanae*, *C. baileyi*, and *Cryptosporidium* sp. chipmunk

Received 11 September 2020; Accepted 11 October 2021; Published 15 December 2021

**Author affiliations:** [1]Center for Emerging and Zoonotic Diseases, College of Veterinary Medicine, South China Agricultural University, Guangzhou 510642, PR China; [2]Guangdong Laboratory for Lingnan Modern Agriculture, Guangzhou 510642, PR China; [3]Division of Foodborne, Waterborne and Environmental Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA.

***Correspondence:** Yaoyu Feng, yyfeng@scau.edu.cn; Lihua Xiao, lxiao1961@gmail.com

**Keywords:** *Cryptosporidium felis*; comparative genomics; codon usage; GC content; reductive evolution.

**Abbreviations:** AOX, alternative oxidase; dUTP, deoxyuridine triphosphate; ENC, effective number of codons; GMP, guanosine monophosphate; MQO, malate quinone oxidoreductase; PCA, principal components analysis; PNO, pyruvate: NADP+ oxidoreductase; RSCU, relative synonymous codon usage; TCA, tricarboxylic acid; TRAP, thrombospondinrelated adhesive protein; XMP, xanthosine 5′-phosphate.

†These authors contributed equally to this work

**Data statement:** All supporting data, code and protocols have been provided within the article or through supplementary data files. Six supplementary tables and eight supplementary figures are available with the online version of this article.

genotype I as well as two gastric *Cryptosporidium* species *C. andersoni* and *C. muris* [3–8]. These whole genome sequence data are available in CryptoDB (https://cryptodb.org/cryptodb/) and NCBI (https://www.ncbi.nlm.nih.gov/assembly/?term=cryptosporidium) databases. All *Cryptosporidium* genomes appear to be approximately 9 Mb organized in eight chromosomes. Nearly complete synteny of genome organization was observed among intestinal *Cryptosporidium* species, while gastric species are more divergent [3, 6].

Results of comparative genomics analyses suggest that the core metabolic pathways of *Cryptosporidium* spp. are highly streamlined [9]. They have lost the ability of *de novo* biosynthesis of most nutrients. In addition, compared with the gastric *Cryptosporidium* species *C. muris* and *C. andersoni*, intestinal species have lost the TCA cycle and conventional oxidative phosphorylation chain [3]. In gene content, they differ from each other in copy numbers of subtelomeric genes encoding several protein families potentially involved in the invasion and host specificity, such as mucin-type glycoproteins (MUC), insulinase-like proteases (INS), and MEDLE proteins [3, 6, 10].

GC content is a core measure of the genomic features of organisms but varies greatly among major groups of apicomplexan parasites. *Cryptosporidium* spp. have relatively low GC contents; except for the 24.3% in *C. baileyi*, the GC content of other *Cryptosporidium* species is nearly 30.0% [6]. In *Plasmodium* spp., the overall genomic GC content is also low, ranging from 19.3–42.3% [11]. Among them, the genomes of *P. vivax* and *P. knowlesi* are less AT-biassed with GC content of 40.5 and 38.6%, respectively. This is especially the case in coding sequences and the third codon position. This variation in genomic GC content has led to differences in codon usage patterns between *P. vivax* and *P. falciparum*, *P. berghei*, *P. chabaudi*, or *P. yoelli*, all of which have extremely low GC content [12]. The diversity in genomic GC content is thought to have important evolutionary and biological significance. It has been suggested that interspecies variation in GC content is a strategy used by both prokaryotes and eukaryotes in their adaption to the diverse environment [13, 14]. Recent evidence suggests that rapid pathogen evolution is responsible for variations in genomic GC content within the genus *Plasmodium* [15].

The availability of whole genome sequence data has remarkably improved our understanding of the metabolic characteristics, genome evolution, and host invasion of *Cryptosporidium* spp. In this study, we sequenced the genome of human-pathogenic *C. felis* and conducted a comparative genomic analysis with other sequenced *Cryptosporidium* species. Results of the analysis have revealed a substantially different codon usage in the *C. felis* genome.

**Impact Statement**

*Cryptosporidium* spp. are apicomplexan parasites infecting the gastrointestinal tract of humans and other vertebrates. Among them, *Cryptosporidium felis* is a host-adapted intestinal *Cryptosporidium* species in cats, but also one of the five most common human-pathogenic species. In this study, we sequenced the genome of *C. felis* for the first time and conducted a comparative genomic analysis. We found that *C. felis* has much higher genomic GC content than other *Cryptosporidium* species, leading to its unique codon usage. A further reduction in metabolism and invasion-related proteins was also noticed, which could contribute to the narrow host range of *C. felis*. Our study documents a notable variation in the genomic GC content among *Cryptosporidium* spp. for the first time, adding to the genetic diversity in this unique group of apicomplexans.

## METHODS

### Specimen processing and genomic sequencing

The *C. felis* isolate 44884 was collected from a human patient in Nebraska, USA and diagnosed to *Cryptosporidium* species by sequence analysis of the small subunit rRNA gene [16]. Oocysts were purified from the specimen using caesium chloride gradient centrifugation and immunomagnetic separation as previously described [17]. The purified *C. felis* oocysts were freeze-thawed five times and digested overnight with protease K. Genomic DNA was extracted from the oocyst suspension using the QIAampDNA Mini Kit (Qiagen Sciences, Maryland, 20874, USA) and amplified using the REPLI-g Midi Kit (Qiagen GmbH, Hilden, Germany). The sequencing of *C. felis* genome was performed on an Illumina HiSeq 2500 after the generation of a 250 bp paired-end library using the Illumina TruSeq (v3) kit (Illumina, San Diego, CA). The CLC Genomics Workbench 11 (https://www.qiagenbioinformatics.com/products/clc-genomics-workbench) was used to trim adapter sequences and sequences with low quality (phred score less than 25), and to assemble *de novo* the genome with a word size of 63 and bulb size of 500. In a secondary analysis, SPAdes 3.1 (http://cab.spbu.ru/software/spades/) was used to assemble the genome with the word size 63 to verify the outcome of the CLC genome assembly.

### Data sources for other *Cryptosporidium* spp. and related apicomplexans

The whole genome sequence and the raw sequence data of other *Cryptosporidium* spp. (*C. hominis* TU502, *C. parvum* IOWA II, *C. meleagridis* UKMEL1, *C. ubiquitum* 39726, *Cryptosporidium* sp. chipmunk genotype I 37763, *C. baileyi* TAMU-09Q1, *C. bovis* 45015, *C. ryanae* 45019, *C. muris* RN66, and *C. andersoni* 30847) were obtained from CryptoDB (https://cryptodb.org/cryptodb/) and NCBI (https://www.ncbi.nlm.nih.gov/sra/?term=Cryptosporidium). The whole genome sequence data of *Plasmodium falciparum* 3D7 and *Toxoplasma gondii* GT1 used in this study were download from PlasmoDB (https://plasmodb.org/plasmo/) and ToxoDB (https://toxodb.org/toxo/), respectively.

### Genome structure analysis

Mauve 2.3.1 [18] and MUMmer 3.2.3 [19] were used to align the assembled *C. felis* genome and genomes of *C. parvum* IOWA II (from CryptoDB), *C. ubiquitum* [3], and *C. andersoni* [3] with default parameters. The completeness of *C. felis* and other *Cryptosporidium* genomes was assessed using BUSCO 4.0.6 [20] through searching the 446 core single-copy orthologs of apicomplexan parasites. The syntenic relationship of genomes was visualized by using Circos 0.69 [21] based on the regions with orthologous genes between *C. felis* and other three genomes. The tRNA genes within the *C. felis* genome were identified by using tRNAscan-SE 2.0.0 [22] and ARAGORN 1.2.36 [23] with the default settings. RNAmmer 1.2 [24] and BLASTN [25] were used to identify the ribosomal RNA genes. Other genomic features were calculated using in-house scripts.

### Analysis of genome rearrangements and gene gains or deletions

Genome alignments of *C. felis* and *C. parvum* IOWA II were constructed by using the progressive alignment algorithm of the Mauve to identify sequence rearrangements in the *C. felis* genome. These sequence rearrangements identified were verified by manual inspection of results of read mapping in the genome assembly for read coverage in the junction areas. Differences in gene numbers in the homologous gene families between *C. felis* and other *Cryptosporidium* species were identified using OrthoMCL [26]. The missing genes in *C. felis* were verified by manual inspection of the Mauve alignments of *C. felis* and *C. parvum* IOWA II genomes. PCR analysis of genomic DNA of *C. felis* was employed to further confirm the rearrangements and gene deletions, using primers spanning the joint of the predicted rearrangements (Table S1, Fig S1, available in the online version of this article) and missing genes (Table S1, Fig S2a). The PCR amplification was performed using a premixed reagent (DreamTaq PCR Master Mix, Fermentas, Vilnius, Lithuania) for 35 cycles of 94 °C for 45 s, 55 °C for 45 s, and 68 °C for 60 s, with an initial denaturation (94 °C for 5 min) and a final extension (72 °C for 7 min). The PCR products were analysed using 1.5% agarose electrophoresis and sequenced in both directions by Sangon Biotech (Shanghai, China) on an ABI3730 autosequencer. The obtained nucleotide sequences were aligned with the target sequences using ClustalX (http://www.clustal.org/).

### Gene prediction and functional annotation

The protein-encoding genes in the *C. felis* genome were predicted using AUGUSTUS 2.7 [27], GeneMark-ES 4.0 [28], SNAP 2013-02-16 [29], and softberry-FGENESH [30] with the default settings. AUGUSTUS and SNAP were trained with the gene model of the published *C. parvum* IOWA II genome before gene prediction. The gene prediction results obtained were combined to reach a final gene set using Evidence Modeller [31] and the outcome was compared with the gene annotation of *C. parvum* IOWA II to ensure that there are no erroneous deletions or fragmentation of genes in each contig. The predicted genes of *C. felis* were annotated using BLASTP 2.7 [25] analysis of the GenBank NR database implemented in Blast2GO [32]. Proteins with GPI anchor sites were identified using online server GPI-SOM [33]. TMHMM 2.0 [34] and SignalP 4.1 [35] were used to predict transmembrane domains and signal peptides of proteins, respectively. The web server KAAS [36] was used to analyse the metabolism of *Cryptosporidium* spp. and other apicomplexans with the eukaryote gene model and the Bi-directional Best Hit method. LAMP (Library of Apicomplexan Metabolic Pathways, release-2) [37], KEGG (Kyoto Encyclopaedia of Genes and Genomes) (https://www.genome.jp/) and Pfam (http://pfam.xfam.org/) [38] were used to further annotate functional proteins, catalytic enzymes, and metabolic pathways.

### Analyses of GC content and codon usage

The GC contents of the genomes and protein-coding genes were calculated using in-house scripts. The cusp mode of EMBOSS explorer was used to calculate the GC content at each of the three codon positions and create a codon usage table to analyse the codon usage frequency (http://www.bioinformatics.nl/emboss-explorer/). DAMBE [39] and CodonW (http://codonw.sourceforge.net) were used to count amino acid frequency, relative synonymous codon usage (RSCU), and the effective number of codons (ENC). Synonymous codons with RSCU values equal to 1.0 represent no codon usage bias for the amino acid. RSCU values <1.0 and >1.0 indicate negative codon usage bias and positive codon usage bias, respectively. RSCU values <0.6 and >1.6 represent under-represented and over-represented codons, respectively [40, 41]. The ENC values range from 20 to 61; a value near 61 means that synonymous codons are used randomly, while a value of 20 indicates that only one codon is used per amino acid [42, 43]. Tbtools was used to perform the principal component analysis (PCA) of the RSCU data (https://github.com/CJ-Chen/Tbtools/).

### Analyses of evolutionary pressure for codon usage

Parity rule 2 (PR2) plots were used to measure the effect of mutation and selection on the codon usage. According to the Chargaff second parity rule (PR2), a double-stranded DNA has the same number of A and T residues, and the same number of G and C residues. In the PR2 plot, if the plot lies on the centre, where both coordinates equal to 0.5 (A=T and G=C), there are no mutational pressure and natural selection for codon usage. In contrast, if the purines and pyrimidines are distributed in the four regions of PR2 plot, the codon usage is affected by these two types of evolutionary pressures [43, 44]. The effective number of codons against the GC content at the third codon position (ENC-GC3) is widely used to assess the effect of base composition on codon usage. If a gene lies onto or near the expected curve of the ENC-GC3 plot, the codon usage is mainly determined by mutational pressure associated with the GC-composition bias. In contrast, if a gene lies below the expected curve, the codon usage is determined by other factors such as natural selection [42]. The neutrality (GC12 vs GC3) analysis is used to evaluate the relationship of GC contents at the three positions on the codon usage. In

**Table 1.** Genomic characteristics of *Cryptosporidium* spp

| Category | Chom | Cpar | Cmel | Cubi | Cchi | Cfel | Cbai | Cbov | Crya | Cmur | Cand |
|---|---|---|---|---|---|---|---|---|---|---|---|
| No. of chromosomes | 8 | 8 | – | – | – | – | – | – | – | – | – |
| Total length of assembly (Mb) | 9.06 | 9.10 | 8.97 | 8.97 | 9.05 | 8.55 | 8.50 | 9.11 | 9.06 | 9.21 | 9.09 |
| No. of super contigs | 53 | 8 | 52 | 39 | 50 | 133 | 153 | 55 | 93 | 42 | 115 |
| GC content (%) | 30.1 | 30.3 | 31.0 | 30.8 | 32.0 | 39.6 | 24.3 | 30.7 | 32.9 | 28.4 | 28.5 |
| No. of genes | 3959 | 3944 | 3753 | 3766 | 3783 | 3775 | 3728 | 3723 | 3711 | 3938 | 3904 |
| Total length of CDS (Mb) | 6.95 | 6.96 | 6.94 | 7.06 | 6.94 | 6.41 | 6.69 | 6.80 | 6.74 | 6.93 | 6.76 |
| GC content in CDS (%) | 31.8 | 31.9 | 32.4 | 33.0 | 33.6 | 40.4 | 25.6 | 31.8 | 33.9 | 30.1 | 30.0 |
| Gene density (genes/Mb) | 437.0 | 433.4 | 418.4 | 419.8 | 418.0 | 441.5 | 438.6 | 408.7 | 409.6 | 425.7 | 429.5 |
| Percentage coding (%) | 76.7 | 76.5 | 77.4 | 78.7 | 76.7 | 75.0 | 78.7 | 74.6 | 74.4 | 74.9 | 74.4 |
| No. of genes with intron | 401 | 437 | 482 | 501 | 515 | 506 | 763 | 571 | 602 | 571 | 489 |
| % genes with introns | 10.0 | 11.1 | 12.8 | 13.3 | 13.6 | 13.4 | 20.5 | 15.3 | 16.2 | 14.5 | 12.5 |
| No. of tRNA | 45 | 45 | 45 | 45 | 45 | 45 | 46 | 45 | 45 | 45 | 44 |
| No. of $tRNA^{Met}$ | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| No. of proteins with signal peptide | 393 | 412 | 394 | 399 | 396 | 387 | 344 | 366 | 329 | 323 | 311 |
| No. of proteins with transmembrane domain | 852 | 870 | 786 | 774 | 793 | 747 | 813 | 781 | 774 | 836 | 843 |
| No. of proteins with GPI-anchor | 58 | 64 | 55 | 50 | 57 | 62 | 57 | 62 | 57 | 52 | 47 |

Chom, *Cryptosporidium hominis*; Cpar, *C. parvum*; Cmel, *C. meleagridis*, Cubi, *C. ubiquitum*; Cchi, *Cryptosporidium* sp. chipmunk genotype I; Cfel, *C. felis*; Cbai, *C. baileyi*; Cbov, *C. bovis*; Crya, *C. ryanae*; Cmur, *C. muris*; Cand, *C. andersoni*.

the neutrality plot, a significant correlation ($P<0.05$) between GC12 and GC3, and the slope of the regression line near 1.0 indicate that the mutational pressure plays a more important role than natural selection in shaping the codon usage [45]. The vhcub [46] was used to draw ENC-GC3 plots, while BCAWT [47] was employed to draw GC12-GC3 plots and PR2 plots.

## Comparative genomics and phylogenetic analysis

OrthoMCL [26] was used to identify the homologous gene families among *Cryptosporidium* species with a e-value threshold of le-8. VennPainter (https://github.com/linguoliang/VennPainter) was used to map shared orthologs and species-specific genes among *C. felis*, *C. parvum*, *C. hominis*, *C. meleagridis*, and *C. ubiquitum*. The MEME suite v5.3.3 [48] was used to identify motif sequences within proteins. A network diagram to show the relationship among proteins in *C. felis*, *C. parvum*, and *C. meleagridis* was visualized using Gephi (https://gephi.org/) with the Fruchterman-Reingold layout. The proteomes of the above species were subjected to BLASTP homologous analysis, and the result was screened (protein pairs sharing 30% identity over 100 amino acids) and used as the input file of the visualization. Pfam was employed to identify and compare the invasion-related proteins and transporter proteins among *C. felis* and other *Cryptosporidium* species. The results of KAAS and online data of LAMP were used in comparative analysis of metabolism among *Cryptosporidium* spp.

## Phylogenetic analysis

The whole sequence reads of *C. parvum*, *C. hominis*, *C. tyzzeri*, *C. meleagridis*, *C. ubiquitum*, *Cryptosporidium* sp. chipmunk genotype I,*C. felis*, *C. baileyi*, *C. bovis*, *C. ryanae*, *C. andersoni*, and *C. muris* were trimmed using Trimmomatic 0.36 [49] to remove adapter sequences and the low-quality bases with an quality below 20. Bowtie 2.3.5.1 [50] was used to map the filtered reads to the *C. parvum* IOWA II genome with default parameters, and the generated bam files of each *Cryptosporidium* species were sorted by samtools 1.12 [51]. The mpileup mode of bcftools 1.10.2 [52] was employed to merge the sorted bam files and generate a bcf files including the insertions or deletions (indels) and the single nucleotide polymorphisms (SNPs) among *Cryptosporidium* species. The call mode of bcftools 1.10.2 was used to generate a vcf file including the SNPs from the bcf file with default parameters and the filter mode was used to discard the SNPs with quality lower than 30. The vcf file including high-quality SNPs was converted to a file in Phy format using a python script vcf2phylip.py (https://github.com/edgardomortiz/vcf2phylip). The SNP-contained Phy file was used to generate the phylogenetic tree among *Cryptosporidium* spp. using the maximum-likelihood (ML) method implemented in RAxML 8 [53]. A ML tree of SKSR protein sequences from *C. parvum*, *C. hominis*, *C. ubiquitum*, *C. meleagridis*, and *C. felis* was constructed using MEGA 6 [54].

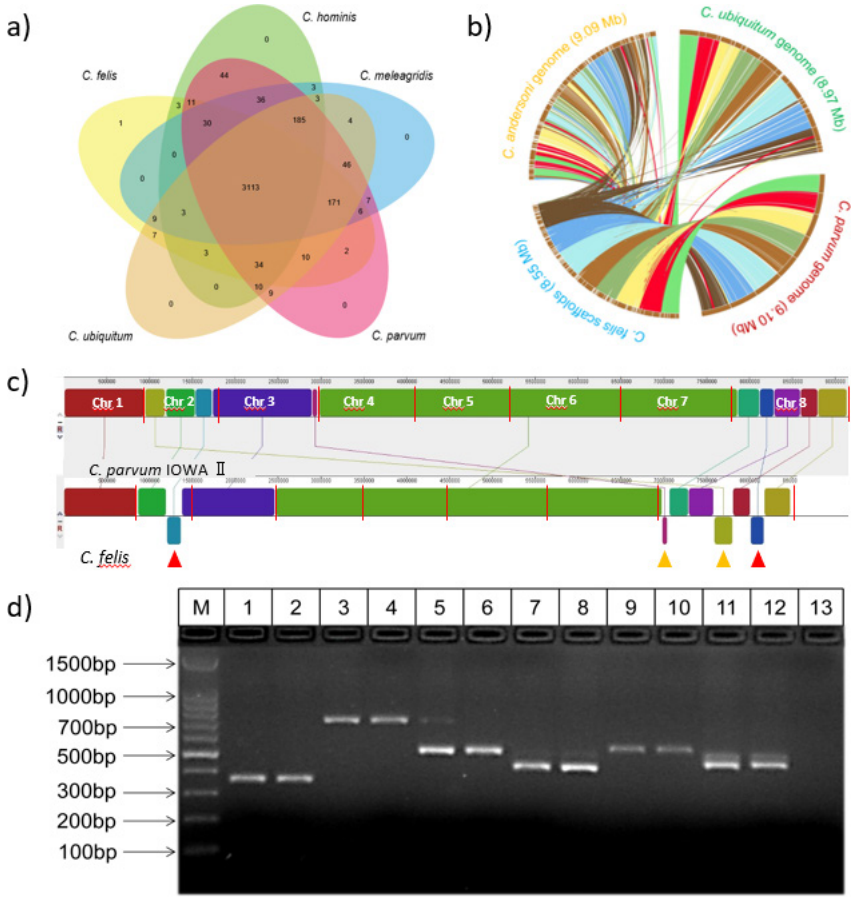


**Fig. 1.** Syntenic relationship of the genomes and shared orthologous genes among several *Cryptosporidium* species. (a) Venn diagram of shared orthologous genes and species-specific genes among *C. felis*, *C. hominis*, *C. parvum*, *C. meleagridis*, and *C. ubiquitum*. (b) Genomic synteny between *C. felis*, *C. parvum*, *C. ubiquitum*, and *C. andersoni*. Syntenic sequences between different genomes are connected with lines. (c) Structural organization of *C. felis* genome comparing to published *C. parvum* IOWA II genome ordered by chromosome (separated from each other by red vertical lines). Sequence rearrangements are present in chromosomes 2 and 8 (red, inversion; yellow, inversion + translocation). The collinear blocks (conserved segments of sequences from the genomes) have the same colours. Assembled contigs of the *C. felis* genome are ordered according to *C. parvum* chromosomes and separated by red vertical lines. (d) Confirmation of sequence inversions and translocations in the *C. felis* genome by PCR. M, 100 bp molecular markers; lanes 1 and 2, sequence inversion and translocation in contig 13 in chromosome 8; lanes 3 and 4, sequence inversion in contig 14 in chromosome 8; lanes 5 and 6, sequence inversion in contig 20 in chromosome 2; lanes 7 and 8, sequence inversion in contig 26 in chromosome 2; lanes 9 and 10, sequence inversion and translocation in contig 35 in chromosome 8; lanes 11 and 12, sequence inversion and translocation in contig 40 in chromosome 8; lane 13, negative control for PCR.

## RESULTS

### Genome sequencing and general genomic features

The genome of a human isolate of *C. felis* 44884 was sequenced using the Illumina sequencing technology, producing 6.9 million 250 bp paired-end reads. After genome assembly and filtering of contigs from bacteria, fungi and the host, a *C. felis* genome of 8.55 Mb in 133 contigs was obtained, with an N50 of 149020 bp at an average 96.7-fold coverage. The completeness of the *C. felis* genome was similar to genomes of other *Cryptosporidium* species in a BUSCO analysis of whole genome sequence data (Fig. S3). Altogether, 3775 protein-encoding genes were predicted from the *C. felis* genome, with gene content and density similar to those of other *Cryptosporidium* species (Table 1). Ortholog analysis indicated that *C. felis* shared most of the protein-encoding genes with other intestinal species such as *C. parvum*, *C. hominis*, *C. meleagridis*, and *C. ubiquitum*, while a group including six *C. felis*-specific SKSR genes (with SK and SR repeats) were identified (Fig. 1a). They formed two gene clusters, one including *Cfel_35.34*, *Cfel_35.5*, *Cfel_35.6*, and *Cfel_35.37* at the 5′ end of chromosome 8 and another including *Cfel_48.7* and *Cfel_136.1* at the 3′ end of chromosome 8. The sequence alignment of the SKSR proteins from different *Cryptosporidium* species showed that the six *C. felis*-specific SKSR proteins have lost two motifs (motifs 1 and 8) and but gained a specific motif (motif 6) (Fig. S4). The genes lost in *C. felis* compared with *C. parvum* are mostly subtolomeric ones, including one cluster of five mucin genes in chromosome 2, two clusters of six MEDLE genes, two insulinase-like protease

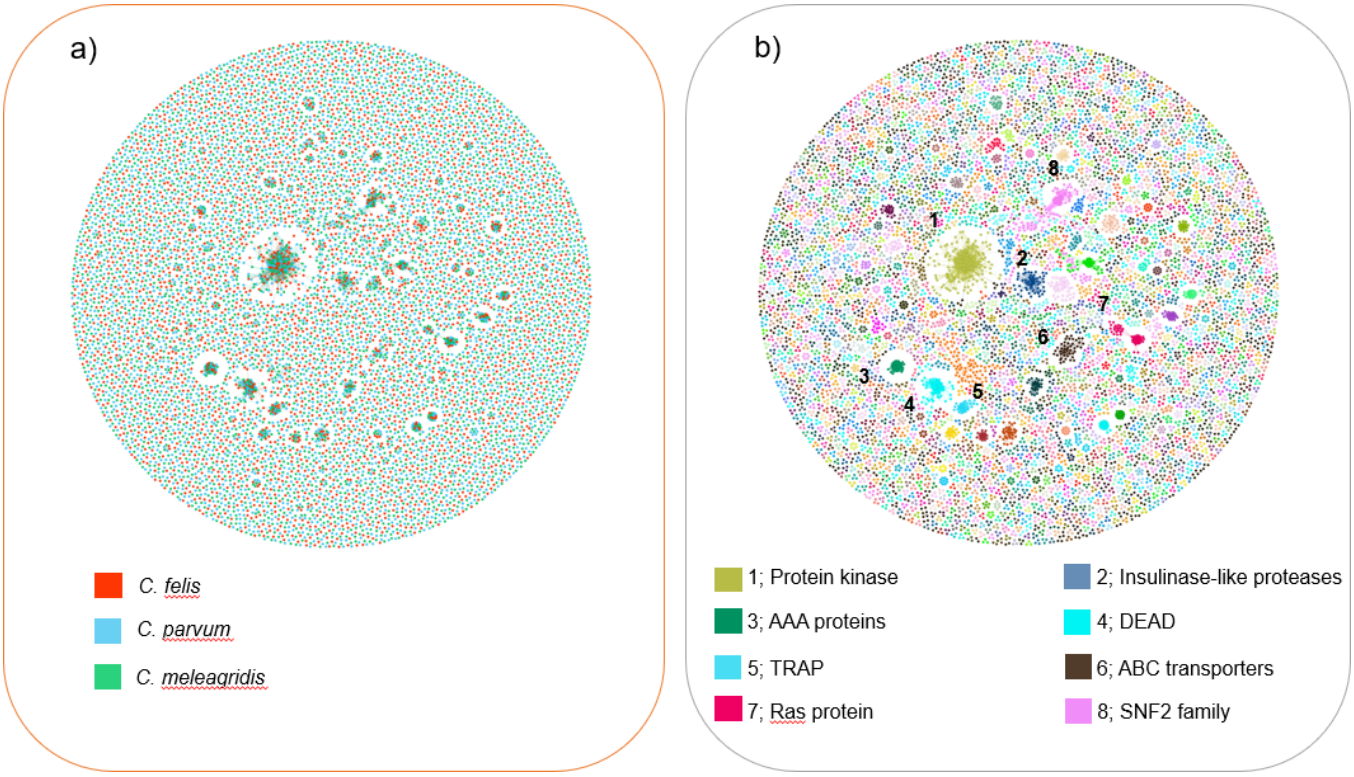


**Fig. 2.** Orthologous clusters of all proteins according to sequence similarity among *Cryptosporidium felis*, *C. parvum*, and *C. meleagridis*. (a) Major cluster of proteins among *C. felis*, *C. parvum* and *C. meleagridis*, which are presented by red, blue, and green dots, respectively. (b) Major protein families in different colours within the *Cryptosporidium* proteomes.

genes in chromosomes 5 and 6, and one cluster of three SKSR genes in chromosome 8.

## Rearrangements in *Cryptosporidium felis* genome

In contrast to gastric species *C. andersoni*, *C. felis* has nearly complete genome synteny with intestinal species *C. parvum* and *C. ubiquitum* (Fig. 1b). However, several genome rearrangements were found in *C. felis* compared with *C. parvum* (Fig. 1c). Sequence inversions were observed in chromosomes 2 (~45.4 kb containing 16 genes in contig 20, and ~126.5 kb containing 52 genes in contig 26) and 8 (~161.7 kb containing 54 genes in contig 14). In addition, both sequence inversions and translocations were observed simultaneously in chromosome 8 (~216.2 kb containing 107 genes in contig 13,~64.4 kb containing 21 genes in contig 35, and ~9 kb containing four genes in contig 40). To confirm these sequence inversions and translocations, we remapped the raw sequencing reads to the junction of the rearranged sequences. The high read coverages in these areas indicates that these rearrangements were not due to assembly errors (Fig. S5). These sequence re-arrangements were also observed in the genome assembly constructed using SPAdes. PCR analyses of the sequences that span over the joints had further confirmed the existence of the sequence inversions and translocations in the *C. felis* genome. The expected PCR products were generated in each of the cases (Fig. 1d).

## Major families of protein-coding genes in *Cryptosporidium* spp

Homolog cluster analysis of the predicted proteomes showed that *C. felis* shares genes encoding major protein families with *C. parvum* and *C. meleagridis* (Fig. 2a). The numbers of genes encoding protein kinases (71 in Cluster 1) and insulinase-like peptidases (21 in Cluster 2) of *C. felis* were slightly less than those of *C. parvum* (79 and 23, respectively) and *C. meleagridis* (78 and 22, respectively). *Cryptosporidium parvum* and *C. meleagridis* have the same number of AAA (Cluster 3), DEAD (Cluster 4) and SNF2 (Cluster 8) encoding genes (25, 39 and 16, respectively), whereas *C. felis* has lost one gene each encoding AAA and DEAD, and two genes encoding SNF2. *Cryptosporidium felis* has 13 genes encoding thrombospondin-related adhesive proteins (TRAPs) (Cluster 5), compared with 12 in *C. parvum* and *C. meleagridis*. *Cryptosporidium parvum* and *C. meleagridis* possess 21 genes encoding ABC transporters (Cluster 6), compared with only 19 in *C. felis*. There are eight, nine, and ten Ras protein-encoding genes (Cluster 7) in *C. felis*, *C. parvum*, and *C. meleagridis*, respectively. (Fig. 2b).

## High GC content in *Cryptosporidium felis* genome

Among the *Cryptosporidium* spp. analysed, *C. felis* has the highest GC content in the genome (39.6%) and coding sequences (40.4%) (Table 1). The GC contents in protein-encoding genes

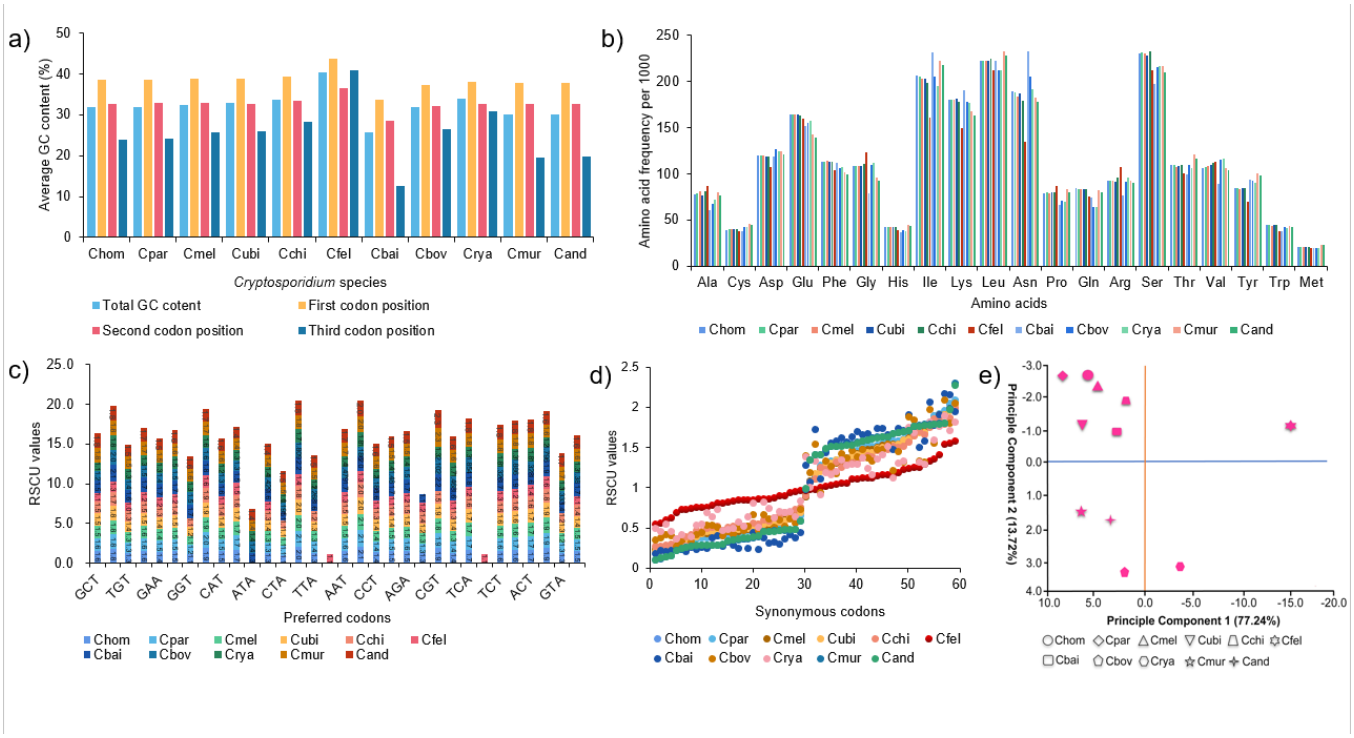


**Fig. 3.** Relationship of GC content and codon usage in *Cryptosporidium* spp. (a) Average GC content of protein-encoding sequences at each of the three codon positions in *Cryptosporidium* spp. (b) Amino acid frequency in *Cryptosporidium* spp. (c) Preferred codons (RSCU values greater than 1.0) used by eleven *Cryptosporidium* species for each amino acid. The RSCU values are shown inside the stacked blocks. (d) Distribution of relative synonymous codon usage (RSCU) values in synonymous codons of eleven *Cryptosporidium* species. No codon usage bias: RSCU values=0; negative codon usage bias: RSCU values <1.0; positive codon usage bias: RSCU values >1.0; over-represented codons: RSCU values >1.6; under-represented codons: RSCU values <0.6. (e) Results of principal component analyses (PCA) of RSCU data from eleven *Cryptosporidium* species. Chom: *Cryptosporidium hominis*; Cpar: *C. parvum*; Cmel: *C. meleagridis*; Cubi: *C. ubiquitum; Cryptosporidium* sp. chipmunk genotype I; Cfel: *C. felis*; Cbai: *C. baileyi*; Cbov: *C. bovis*; Crya: *C. ryanae*; Cmur, *C. muris*; Cand: *C. andersoni*.

varied mostly from 30–50% in *C. felis* compared with 10–40% in other *Cryptosporidium* species (Fig. S6). The GC contents in the first codon (GC1), second codon (GC2) and third codon (GC3) positions in *C. felis* were 44.0, 36.2 and 40.9%, respectively, compared with 37.8–38.8%, 32.6–32.9% and 19.4–26.0% in other *Cryptosporidium* species (Fig. 3a). The much higher GC content of GC3 in *C. felis* indicates that the difference in the GC content between *C. felis* and other *Cryptosporidium* spp. mainly comes from this position.

## Codon usage pattern in *Cryptosporidium felis*

Genomic GC content impacts amino acid usage of *Cryptosporidium* spp. While other *Cryptosporidium* species exhibit high usages of Asn, Asp, Lys, Glu, Ile, Ser, and Leu residues encoded by AT-rich codons, *C. felis* prefers Ala, Gly, Arg, and Pro residues encoded by GC-rich codons (Fig. 3b). To further explore the impact of differences in the GC content, RSCU was calculated to compare the preferred codons (RSCU value >1.0) among *Cryptosporidium* spp. Although all *Cryptosporidium* species prefer to use codons ending with A/T, two preferred codons ending with G/C (TTG and TCC) were found in *C. felis* (Fig. 3c, Table S2). In addition, the under-represented codons (RSCU value <0.6) are generally ending with G/C and the over-represented codons (RSCU value >1.6) with A/T in other *Cryptosporidium* spp. However, such under-represented and over-represented codons are obviously fewer in *C. felis*, and the RSCU values for most codons are between 0.6–1.0 and all these codons are ending with G/C (Fig. 3d, Table S2). This difference in codon usage was also observed in the genes encoding the invasion-related insulinase-like protease, mucin-type glycoprotein, and thrombospondin-related adhesive protein families between *C. felis* and other *Cryptosporidium* species (Fig. S7). The result of a PCA of the RSCU data showed an obvious separation of *C. felis* from other *Cryptosporidium* species (Fig. 3e).

There are 45 tRNA genes in genomes of *Cryptosporidium* spp. except for *C. baileyi*, which uses an additional tRNA of TTG for glutamine (46), and *C. andersoni*, which does not have a tRNA of CCA for proline (44) (Table 1). *Cryptosporidium felis* has the same number of the tRNA genes as most other *Cryptosporidium* species but uses a tRNA for GTG instead of GTA for valine (Table S2).

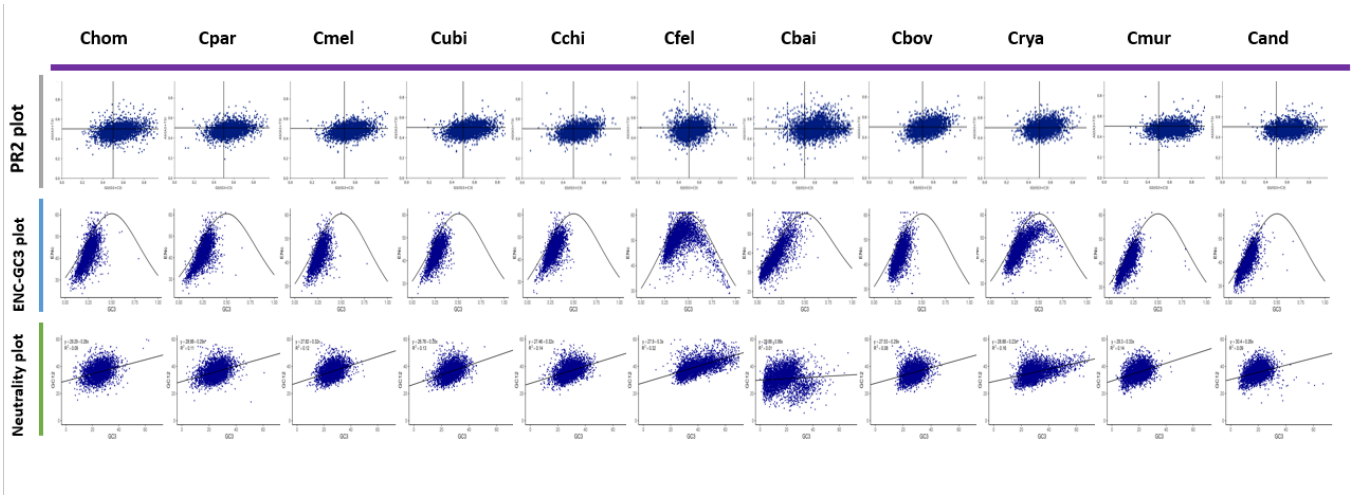


**Fig. 4.** Evolutionary pressure associated with codon usage in *Cryptosporidium* spp. The parity Rule 2 (PR2) plots show ratios of A3/(A3 +T3) against G3/(G3 +C3). Both coordinates equal to 0.5 in the centre of the plot denote there is no bias of mutation or selection in this gene. The plots of effective number of codons (ENC)-the GC content at the third codon position (GC3) show the relationship between ENC values and GC3. The black expected curve is the expected ENC against GC3. The neutrality (GC12-GC3) plots show the correlation between GC12 (average GC content at first and second positions of the codon) and GC3. The linear regression of GC12 against GC3 is indicated by the black line. Chom: *Cryptosporidium hominis*; Cpar: *C. parvum*; Cmel: *C. meleagridis*; Cubi: *C. ubiquitum; Cryptosporidium* sp. chipmunk genotype I; Cfel: *C. felis*; Cbai: *C. baileyi*; Cbov: *C. bovis*; Crya: *C. ryanae*; Cmur, *C. muris*; Cand: *C. andersoni*.

## Natural selection shapes codon usage in *Cryptosporidium felis*

The role of mutational pressure and natural selection in shaping the codon usage of *Cryptosporidium* spp. was assessed using PR2, ENC-GC3, and neutrality plots (Fig. 4). In the PR2 plots, T and G are used more frequently than A and C in most *Cryptosporidium* spp. The imbalanced usage of A+T and G+C suggested that both mutational and natural selection impact the codon usage of *Cryptosporidium* spp. In contrast, *C. felis* has a more balanced GC usage, with most genes located in the centre of the PR2 plot. Results of the ENC-GC3 analysis supported the more prominent role of natural selection in shaping codon usage in *C. felis*. While many genes in other *Cryptosporidium* spp. lie below the expected curve rather than on or around the curve, far more genes are doing so in *C. felis*, indicating that natural selection rather than mutational pressure is more important in shaping the codon usage in *C. felis*. Moreover, the neutrality plot analysis revealed that the GC12 is more positively correlated with GC3 in *C. felis* ($R^2$=0.32; $P$=1.16×$10^{-320}$) than in other species ($R^2$:0.01–0.16; $P$: 4.80×$10^{-145}$-10.29×$10^{-06}$). As the slope of the regression line in *C. felis* (0.30) was far from 1.0, this result further indicated that the natural selection (70%) plays more important role than mutational pressure (30.0%) in shaping the codon usage of the genome.

## Metabolism in *Cryptosporidium felis*

In carbohydrate and energy metabolism, like other intestinal *Cryptosporidium* species, *C. felis* has lost genes encoding core enzymes of the tricarboxylic acid (TCA) cycle and mainly obtains energy through the glycolytic pathway (Fig. 5; Table S3). *Cryptosporidium parvum*, *C. hominis, C. meleagridis*, and *Cryptosporidium* sp. chipmunk genotype I have a cyanide-insensitive alternative oxidase (AOX), which is lost in *C. felis*, *C. ubiquitum*, *C. baileyi*, *C. bovis*, and *C. ryanae* together with the associated ubiquinone biosynthesis pathway (Fig. 5, Table S3). Compared with nine mitochondrial carrier proteins in *C. parvum* and *C. hominis* and eight in *C. meleagridis* and chipmunk genotype I, only seven, six, five, four and three were detected in *C. felis, C. baileyi*, *C. ubiquitum*, *C. ryanae*, and *C. bovis*, respectively (Table S4). All *Cryptosporidium* species must salvage nucleotides from the host via the nucleoside transporter due to the lack of capacity for *de novo* biosynthesis of purines and pyrimidines.

One gene encoding the guanosine monophosphate (GMP) synthetase (ortholog of *cgd5_4520* in *C. parvum, Chro.50499* in *C. hominis*, and *ctyz_00002695* in *C. tyzzeri*) is lost in *C. felis*, indicating that *C. felis* has lost the ability to convert xanthosine 5′-phosphate (XMP) to GMP. The function of deoxyuridine triphosphate (dUTP) diphosphatase is to reduce the accumulation of dUTP and prevent the incorporation of atypical nucleotides into DNA. There are two copies of the gene encoding the enzyme in *C. parvum*, *C. hominis*, and *C. ubiquitum*, but only one in *C. felis*, *C. meleagridis*, *Cryptosporidium* sp. chipmunk genotype I, *C. baileyi*, *C. muris*, and *C. andersoni*.

## Characteristics of invasion-related proteins in *Cryptosporidium felis*

*Cryptosporidium felis* and other intestinal *Cryptosporidium* species possess some unique invasion-related proteins (Fig. S8), with the members of these proteins being variable among species

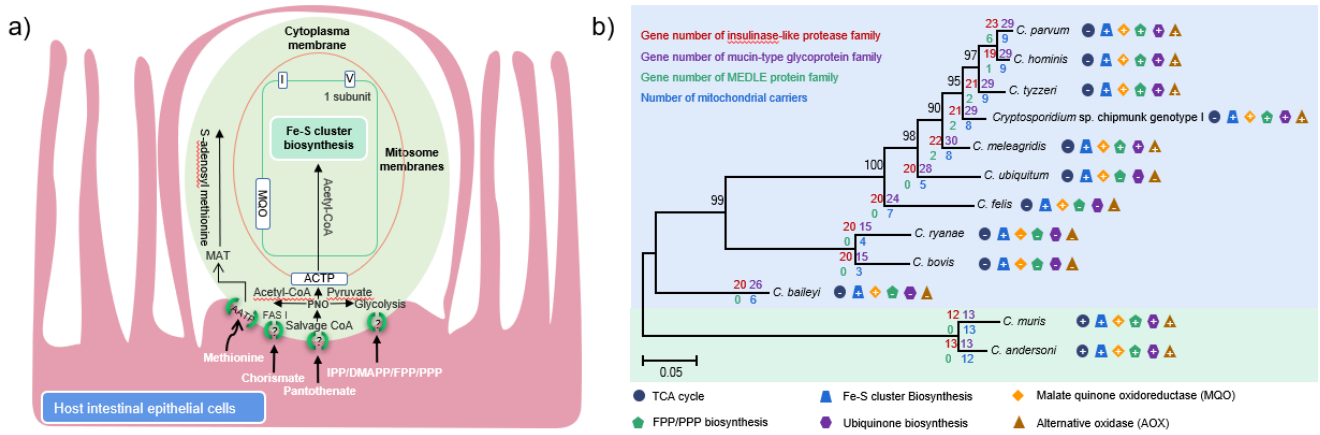


**Fig. 5.** Evolution of mitochondrial metabolism and invasion-related proteins in *Cryptosporidium* spp. (a) Schematic diagram of mitochondrial metabolism in *Cryptosporidium felis*. MAT: methionine adenosyl transferase; AATP: amino acids transporter protein; ACTP: acetyl-CoA transporter protein; PNO: pyruvate:NADP(+) oxidoreductase; CoA: coenzyme A; MQO: malate:quinone oxidoreductase; IPP: isopentenyl diphosphate; DMAPP: dimethylallyl diphosphate; FPP: farnesyl diphosphate; PPP: polyprenyl diphosphate. (b) Reductive evolution in invasion-related proteins, mitochondrial carriers, and metabolic pathways/enzymes during the evolution of *Cryptosporidium* species. The numbers of invasion-related insulinase-like proteases, mucin-type glycoproteins, MEDLE proteins families and mitochondrial carriers in different *Cryptosporidium* species are shown on the branches with different colours. The metabolic pathways/enzymes in *Cryptosporidium* spp. are shown by different shapes after the species name. A plus symbol in the shape denotes the presence of the pathway/enzyme, while a minus symbol denotes the absence of the pathway/enzyme.

(Table S5). Among subtelomeric genes encoding insulinase-like proteases, like most other *Cryptosporidium* species, *C. felis* has lost the *C. parvum*-specific subtelomeric gene encoding INS19-20 in chromosome 6. *Cryptosporidium felis* has further lost five subtelomeric genes encoding mucin glycoproteins (MUC3-7) in chromosome 2, and all six subtelomeric genes encoding MEDLE family proteins in chromosomes 5 and 6. The SNP-based phylogenetic tree showed consistency between losses of invasion-related proteins or metabolic enzymes and the evolutionary relationship among *Cryptosporidium* species. A progressive reduction in the number of these proteins and enzymes was seen with the increased divergence of the intestinal *Cryptosporidium* species from *C. parvum* (Fig. 5b).

### Losses of other genes in *Cryptosporidium felis*

*C. felis* has lost some genes in subtelomeric regions across the eight chromosomes compared with other *Cryptosporidium* spp. (Table S6). The missing genes mostly encoding hypothetical proteins and one-third of them contain signal peptides in *C. parvum* (*cgd1_120*, *cgd1_140*, *cgd2_390*, *cgd2_400*, *cgd2_420*, *cgd2_450*, *cgd3_1750*, *cgd4_10*, *cgd4_3650*, *cgd4_4500*, *cgd5_4520*, *cgd6_5470*, *cgd6_5480*, *cgd6_5490*, *cgd7_1210*, *cgd7_4430*). The genes deletions in *C. felis* were observed at the well-assembled syntenic regions shared with *C. parvum* genome. The absence of some genes in *C. felis*, such as the ortholog of *cgd5_4520* encoding a GMP synthase and orthologs of *cgd6_5480* and *cgd6_5490* encoding two MEDLE proteins had resulted in smaller PCR products for *C. felis* than *C. parvum* (Fig. S2), supporting that these orthologous genes of *C. parvum* were absent in *C. felis*.

## DISCUSSION

Results of the present study have shown a high similarity in genome organization of *C. felis* to other intestinal *Cryptosporidium* species. The size of the *C. felis* genome (8.55 Mb) is slightly smaller than other intestinal *Cryptosporidium* spp. but similar to that of *C. baileyi* (8.50 Mb). Like *C. baileyi*, *C. felis* has fewer genes than other human-pathogenic *Cryptosporidium* species. Nevertheless, genes in the *C. felis* genome are organized in almost complete synteny with other intestinal *Cryptosporidium* spp [3, 6, 10]. Minor sequence inversions and translocations were found in the subtelomeric regions of *C. felis* compared with *C. parvum*. Syntenic breaks in the subtelomeric regions of apicomplexans have been associated with variations in species-specific genes and multigene families [55]. In agreement with observations in the present study, a recent study has shown that sequence rearrangements in the subtelomeric regions of chromosomes have resulted in the loss of some genes encoding secreted proteins in *C. bovis* [8].

A major genomic difference between the *C. felis* and other *Cryptosporidium* spp. is the GC content. *Cryptosporidium felis* has a much higher GC content (39.6%) than other *Cryptosporidium* spp. (24.3–32.0%). This is largely attributed to the high GC content at the third codon position. Variations in

genomic GC content contribute to the differences in codon usage among organisms [56, 57]. Substantial inter-species differences in genomic GC content are rare in apicomplexan parasites. The only known example of such differences within this phylum is the genus *Plasmodium*. While most *Plasmodium* species have A/T-rich genomes, *P. vivax* and *P. knowlesi* have much higher GC content, leading to different codon usage [12]. Mutational pressure and natural selection for certain functional protein classes were suggested to be the causes for variations in genomic GC content among *Plasmodium* spp [58].

As expected, the higher GC content in the *C. felis* genome has led to different codon usage. The nucleotide composition of genomes has an important effect on codon usage [59]. Because of the higher GC content, *C. felis* tends to use more amino acid residues encoded by GC-rich codons than other *Cryptosporidium* spp. Moreover, the codon usage is polarized in AT-rich *Cryptosporidium* species, reflected by the dominance of over-represented codons (RSCU value >1.6) ending with A/T and under-represented codons (RSCU value <0.6) ending with G/C. In contrast, *C. felis* has mostly G/C-ended codons with RSCU values ranging from 0.6 to 1.0. These results suggest that *C. felis* is evolving towards a more balanced G/C codon usage.

The unique codon usage in *C. felis* appears to be driven by natural selection. Mutational pressure and natural selection are two major evolutionary forces driving the evolution of codon usage [60]. A previous study showed that both mutational pressure and natural selection contributed to shaping the codon usage patterns in GC-poor *P. falciparum* and GC-rich *P. vivax* and *P. knowlesi* [61]. It appears that election pressure is a dominant force driving the evolution of codon usage bias in *P. vivax* and *P. knowlesi* while mutational pressure helps to shape the codon usage in *P. falciparum* [62]. In the present study, the results of PR2 analysis showed that mutational pressure and natural selection both affected the codon usage in *C. felis*. In addition, the GC content of genomes constrains ENC and the combination of the two helps to shape the codon usage. The results of the ENC-GC analysis, however, revealed that natural selection plays a more important role in shaping the codon usage of *C. felis*. This was further supported by the results of the neutrality analysis [45].

The GC-related differences in codon usage could potentially affect gene expression in *C. felis*. Although we have limited knowledge of the biological implication of variations in genomic GC content and codon usage in apicomplexan parasites, a recent study has shown that the highly expressed genes in developmental stages of *P. falciparum* prefer to use amino acid residues (Gly, Arg, Ala, and Pro) encoded by GC-rich codons [63]. This is because the energy consumption of biosynthesis is different among amino acids [64]; the biosynthesis of some of the preferred amino acids such as Gly and Ala consume less energy. Therefore, *P. falciparum* maintains high GC content in highly expressed genes. Similarly, *C. felis* prefers less energetically costly Gly and Ala residues encoded by GC-rich codons. As *Cryptosporidium* spp. acquire amino acids from the host instead of biosynthesizing them *de novo* [65], the preferential usage of less energetically costly amino acid residues in *C. felis* could lead to a more harmonious host-parasite relationship. RNA sequencing and comparative transcriptomics analysis are needed to identify GC-content associated differences in gene expression between *C. felis* and other *Cryptosporidium* species.

The progressive reductive evolution in metabolism and invasion-related proteins reflects the host-adapted characteristics of *C. felis*. As reported in other apicomplexan parasites, variation in metabolism among *Cryptosporidium* spp. is probably the result of lineage-specific host adaptation by parasites [66]. In contrast to other intestinal *Cryptosporidium* species, *C. felis* together with *C. ubiquitum*, *C. baileyi*, *C. bovis* and *C. ryanae* [3, 6, 8] has lost the corresponding enzymes involved in both conventional and alternative electron transport systems. The more streamlined metabolism in *C. felis* could be the result of its advanced adaptation to the host environment. Accompanying the reductive evolution in metabolism, *C. felis* has lost some genes encoding invasion-related proteins of *C. parvum*. The loss of these proteins could be responsible for the narrow host range of *C. felis* [3].

The present study has detected some losses of subtelomeric genes in *C. felis*. As reported in *Toxoplasma gondii*, variations in the number of secretory proteins could be responsible for the host range among different strains [67]. In *Cryptosporidium* spp., absences of some genes encoding secretory proteins were detected in *Cryptosporidium* sp. chipmunk genotype I, *C. bovis* and *C. ryanae*, which have narrow host ranges [6, 8]. Similar genes losses were found in the *C. felis* genome. As the duplicated MEDLE genes could not be detected by the current sequencing technology generated in this work, further work generating long sequence reads to obtain a complete, free of gaps genome assembly is needed to validate if MEDLEs are indeed not present in *C. felis*. These additional genes losses could further contribute to the host-adapted nature of *C. felis*.

In conclusion, we obtained whole-genome sequence data from *C. felis* for the first time. Results of comparative genomic analyses indicate that although *C. felis* shares genomic characteristics with other *Cryptosporidium* species, the notably higher GC content, especially at the third codon position, enables *C. felis* to use less energetically costly amino acid residues encoded by GC-rich codons, leading to better adaption to the host environment. The increased GC content appears to be driven mainly by natural selection. This better host adaptation is strengthened by further reductive evolution of metabolism and host invasion-related proteins, probably leading to the feline adapted nature of *C. felis*.

Funding information
This work was supported in part by the Guangdong Major Project of Basic and Applied Basic Research (2020B0301030007), National Natural Science Foundation of China (31820103014, U1901208), 111 Project (D20008), and Innovation Team Project of Guangdong Universities (2019KCXTD001).

Acknowledgements
The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Author contributions
Conceptualization: L.X. and Y.F.; investigation and resources: L.X. and Y.F.; methodology: J.L., Y.G., and L.X.; formal analysis: J.L. and L.X.; writing – original draft preparation: J.L. and Y.G.; writing – review and editing: L.X. and Y.F.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
This research was reviewed and approved by the Ethics Committee of the South China Agricultural University. The faecal specimen was collected with the permission of the patient as part of the routine clinical diagnosis. The study protocol was approved by the institutional review board of the Centres for Disease Control and Prevention and Prevention, USA (No. 990115).

References

1. Feng Y, Ryan UM, Xiao L. Genetic diversity and population structure of *Cryptosporidium*. *Trends Parasitol* 2018;34:997–1011.
2. Xiao L, Feng Y. Zoonotic cryptosporidiosis. *FEMS Immunol Med Microbiol* 2008;52:309–323.
3. Liu S, Roellig DM, Guo Y, Li N, Frace MA, *et al*. Evolution of mitosome metabolism and invasion-related proteins in *Cryptosporidium*. *BMC genomics* 2016;17:1006.
4. Ifeonu OO, Chibucos MC, Orvis J, Su Q, Elwin K, *et al*. Annotated draft genome sequences of three species of *Cryptosporidium*: *Cryptosporidium meleagridis* isolate UKMEL1, *C. baileyi* isolate TAMU-09Q1 and *C. hominis* isolates TU502_2012 and UKH1. *Pathog Dis* 2016;74:ftw080.
5. Sateriale A, Šlapeta J, Baptista R, Engiles JB, Gullicksrud JA, *et al*. A genetically tractable, natural mouse model of cryptosporidiosis offers insights into host protective immunity. *Cell Host Microbe* 2019;26:135–146.
6. Xu Z, Guo Y, Roellig DM, Feng Y, Xiao L. Comparative analysis reveals conservation in genome organization among intestinal *Cryptosporidium* species and sequence divergence in potential secreted pathogenesis determinants among major human-infecting species. *BMC Genomics* 2019;20:406.
7. Nader JL, Mathers TC, Ward BJ, Pachebat JA, Swain MT, *et al*. Evolutionary genomics of anthroponosis in *Cryptosporidium*. *Nat Microbiol* 2019;4:826–836.
8. Xu Z, Li N, Guo Y, Feng Y, Xiao L. Comparative genomic analysis of three intestinal species reveals reductions in secreted pathogenesis determinants in bovine-specific and non-pathogenic *Cryptosporidium* species. *Microb Genom* 2020;6:e000379.
9. Zhu G, Guo F. *Cryptosporidium* metabolism. In: Cacciò SM and Widmer G (eds). *Cryptosporidium: Parasite and Disease*. Vienna: Springer; 2014. pp. 361–379.
10. Guo Y, Tang K, Rowe LA, Li N, Roellig DM, *et al*. Comparative genomic analysis reveals occurrence of genetic recombination in virulent Cryptosporidium hominis subtypes and telomeric gene duplications in Cryptosporidium parvum. *BMC Genomics* 2015;16:320.
11. Videvall E. *Plasmodium parasites* of birds have the most AT-rich genes of eukaryotes. *Microb Genom* 2018;4:e000150.
12. Yadav MK, Swati D. Comparative genome analysis of six malarial parasites using codon usage bias based tools. *Bioinformation* 2012;8:1230–1239.
13. Romiguier J, Ranwez V, Douzery EJP, Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res* 2010;20:1001–1009.
14. Šmarda P, Bureš P, Horová L, Leitch IJ, Mucina L, *et al*. Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc Natl Acad Sci U S A* 2014;111:E4096-102.
15. Nikbakht H, Xia X, Hickey DA. The evolution of genomic GC content undergoes a rapid reversal within the genus Plasmodium. *Genome* 2014;57:507–511.
16. Xiao L, Morgan UM, Limor J, Escalante A, Arrowood M, *et al*. Genetic diversity within *Cryptosporidium parvum* and related *Cryptosporidium* species. *Appl Environ Microbiol* 1999;65:3386–3391.
17. Guo Y, Li N, Lysén C, Frace M, Tang K, *et al*. Isolation and enrichment of *Cryptosporidium* DNA and verification of DNA purity for whole-genome sequencing. *J Clin Microbiol* 2015;53:641–647.
18. Darling AE, Mau B, Perna NT. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 2010;5:e11147.
19. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, *et al*. Versatile and open software for comparing large genomes. *Genome Biol* 2004;5:R12.
20. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31:3210–3212.
21. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res* 2009;19:1639–1645.
22. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997;25:955–964.
23. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* 2004;32:11–16.
24. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007;35:3100–3108.
25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
26. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;13:2178–2189.
27. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* 2004;32:W309-12.
28. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* 2005;33:6494–6506.
29. Korf I. Gene finding in novel genomes. *BMC Bioinformatics* 2004;5:59.
30. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol* 2006;7:S10.
31. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol* 2008;9.
32. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005;21:3674–3676.
33. Fankhauser N, Mäser P. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics* 2005;21:1846–1852.
34. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001;305:567–580.
35. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 2011;8:785–786.
36. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 2007;35:W182-5.
37. Shanmugasundram A, Gonzalez-Galarza FF, Wastling JM, Vasieva O, Jones AR. Library of apicomplexan metabolic pathways:

a manually curated database for metabolic pathways of apicomplexan parasites. *Nucleic Acids Res* 2013;41:D706-13.

38. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, *et al*. Pfam: the protein families database. *Nucleic Acids Res* 2014;42:D222-30.
39. Xia X. DAMBE6: new tools for microbial genomics, phylogenetics, and molecular evolution. *J Hered* 2017;108:431–437.
40. Wong EHM, Smith DK, Rabadan R, Peiris M, Poon LLM. Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. *BMC Evol Biol* 2010;10:253.
41. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* 1986;24:28–38.
42. Wright F. The "effective number of codons" used in a gene. *Gene* 1990;87:23–29.
43. Comeron JM, Aguadé M. An evaluation of measures of synonymous codon usage bias. *J Mol Evol* 1998;47:268–274.
44. Rapoport AE, Trifonov EN. Compensatory nature of Chargaff's second parity rule. *J Biomol Struct Dyn* 2013;31:1324–1336.
45. Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res* 2003;92:1–7.
46. Anwar AM, Soudy M, Mohamed R. Vhcub: virus-host codon usage co-adaptation analysis. *F1000Res* 2019;8:2137.
47. Anwar A. BCAWT: automated tool for codon usage bias analysis for molecular evolution. *JOSS* 2019;4:1500.
48. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;37:W202-8.
49. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
50. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
51. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
52. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, *et al*. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* 2016;32:1749–1751.
53. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
54. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 2013;30:2725–2729.
55. DeBarry JD, Kissinger JC. Jumbled genomes: missing apicomplexan synteny. *Mol Biol Evol* 2011;28:2855–2871.
56. Birdsell JA. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol* 2002;19:1181–1197.
57. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* 2009;10:285–311.
58. Castillo AI, Nelson ADL, Lyons E. Tail wags the dog? Functional gene classes driving genome-wide GC content in *Plasmodium* spp. *Genome Biol Evol* 2019;11:497–507.
59. Hershberg R, Petrov DA. General rules for optimal codon choice. *PLoS Genet* 2009;5:e1000556.
60. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 1991;129:897–907.
61. Bunnik EM, Chung D-W, Hamilton M, Ponts N, Saraf A, *et al*. Polysome profiling reveals translational control of gene expression in the human malaria parasite *Plasmodium falciparum*. *Genome Biol* 2013;14.
62. Gajbhiye S, Patra PK, Yadav MK. New insights into the factors affecting synonymous codon usage in human infecting *Plasmodium* species. *Acta Trop* 2017;176:29–33.
63. Chanda I, Pan A, Dutta C. Proteome composition in *Plasmodium falciparum*: higher usage of GC-rich nonsynonymous codons in highly expressed genes. *J Mol Evol* 2005;61:513–523.
64. Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and Bacillus subtilis. *Proc Natl Acad Sci U S A* 2002;99:3695–3700.
65. Rider SD, Zhu G. *Cryptosporidium*: genomic and biochemical features. *Exp Parasitol* 2010;124:2–9.
66. Song C, Chiasson MA, Nursimulu N, Hung SS, Wasmuth J, *et al*. Metabolic reconstruction identifies strain-specific regulation of virulence in *Toxoplasma gondii*. *Mol Syst Biol* 2013;9:708.
67. Lorenzi H, Khan A, Behnke MS, Namasivayam S, Swapna LS, *et al*. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic *Toxoplasma gondii* genomes. *Nat Commun* 2016;7:10147.