Title: Evolutionary genomics of anthroponosis in *Cryptosporidium*

Authors and Affiliations

Johanna L. Nader[1,2], Thomas C. Mathers[3], Ben J. Ward[3,4], Justin A. Pachebat[5], Martin T. Swain[5], Guy Robinson[6,7], Rachel M. Chalmers[6,7], Paul R. Hunter[1], Cock van Oosterhout[4*], Kevin M. Tyler[1*]

[1]Biomedical Research Centre, Norwich Medical School, University of East Anglia, Norwich, United Kingdom
[2]Department of Genetics and Bioinformatics, Division of Health Data and Digitalisation, Norwegian Institute of Public Health, Oslo, Norway
[3]Earlham Institute, Norwich Research Park, Norwich, United Kingdom
[4]School of Environmental Sciences, Norwich Research Park, University of East Anglia, United Kingdom
[5]Institute of Biological, Environmental & Rural Sciences, Aberystwyth University, Aberystwyth, United Kingdom
[6]*Cryptosporidium* Reference Unit, Public Health Wales Microbiology, Singleton Hospital, Swansea, United Kingdom
[7]Swansea University Medical School, Singleton Park, Swansea, United Kingdom

*Contributed equally to the work

Corresponding authors:
Email: johanna.nader@fhi.no
Telephone: +47 41221727
Address: Norwegian Institute of Public Health, Postbox 4404, Nydalen 0403, Oslo, Norway

Email: c.van-oosterhout@uea.ac.uk
Telephone: +44 1603 592921
Address: School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK

Abstract

Human cryptosporidiosis is the leading protozoan cause of diarrhoeal mortality worldwide, and a preponderance of infections is caused by *Cryptosporidium hominis* and *C. parvum*. Both species consist of several subtypes with distinct geographic distributions and host preferences (i.e. generalist zoonotic and specialist anthroponotic subtypes). The evolutionary processes driving the adaptation to human host, and the population structure remain unknown. In this study, we analyse 21 whole genome sequences to elucidate the evolution of anthroponosis. We show that *C. parvum* splits into two subclades, and that the specialist anthroponotic subtype IIc-a shares a subset of loci with *C. hominis* that are undergoing rapid convergent evolution driven by positive selection. Subtype IIc-a also has an elevated level of insertion-deletion (indel) mutations in the peri-telomeric genes, which is characteristic also for other specialist subtypes. Genetic exchange between subtypes plays a prominent role throughout the evolution of *Cryptosporidium*. Interestingly, recombinant regions are enriched for positively selected genes and potential virulence factors, which indicates adaptive introgression. Analysis of 467 gp60 sequences collected across the world shows that the population genetic structure differs markedly between the main zoonotic subtype (isolation-by-distance) and the anthroponotic subtype (admixed population structure). Finally, we show that introgression between the four anthroponotic *Cryptosporidium* subtypes and species included in this study has occurred recently, probably within the past millennium.

Introduction

Diarrhoeal pathogens cause more mortality than malaria, measles, and AIDS combined[1] and globally, for children under five, *Cryptosporidium* is the leading, vaccine non-preventable cause of diarrhoeal morbidity and mortality[2]. The zoonotic *Cryptosporidium parvum* and the anthroponotic *Cryptosporidium hominis* account for a vast majority of such cases. *C. hominis* and *C. parvum* have consistently been reported as exhibiting a high average global consensus of ~95-97% nucleotide identities[3,4]; yet, the genetic basis for the difference in host range has remained unexplained, and our understanding of host adaptation is confounded by the existence of anthroponotic *C. parvum* isolates (Supplementary Fig. S1). The relatively high level of genomic conservation between these species could be explained by similarity in selection pressures experienced by these parasites that is irrespective of their hosts. For example, *Plasmodium berghei* requires two-thirds of genes for optimal growth during a single stage of its complex life cycle[5]. Alternatively, hybridization amongst isolates of *Cryptosporidium* species could lead to genetic introgression that homogenizes sequence variation. For example, some "generalist" plant pathogens such as the oomycete *Albugo candida* have a huge host range consisting of hundreds of plant species that are parasitized by host-specific subtypes[6]. This pathogen suppresses the immune response of the host plant, enabling hybridization between different subtypes leading to genetic introgression that is thought to fuel the coevolutionary arms race[3,8] . Similarly, in the mosaic-like *Toxoplasma gondii* genomes there are conserved chromosomal haploblocks which are shared across otherwise diverged clades[7].

The ~9.14Mbp *Cryptosporidium* genome comprises 8 chromosomes ranging in size from 0.88 to 1.34Mbp, and has a highly compact coding sequence composition (73.2-77.6%)[8]. Genomic comparisons between the original *C. parvum* Iowa[9] and *C. hominis* TU502[10] reference genomes currently provide an overview of chromosome-wide hotspots for single nucleotide polymorphisms (SNPs), selective pressures, and species-specific genes and duplication events[4,11]. These studies revealed peri-telomeric clustering of hyper-polymorphism and identified several putative virulence factors. Attempts to correlate

genomic changes with phenotypic expression identified only a few shared SNPs between the anthroponotic *C. parvum* and *C. hominis*[12]. Whole genome comparisons found genome-wide incongruence and significant sequence insertion and deletion (indels) events between *C. hominis* and *C. parvum*[13], and recombination at the hypervariable gp60 subtyping locus[14]. Expanding cross-comparisons to include multiple whole genome sequences (WGS) across a range of anthroponotic and zoonotic *C. parvum* and *C. hominis* strains will help to explore these phenotype-associated features, and understand the evolution of human-infective strains.

Here, we have conducted a phylogenetic comparison of 21 WGS, including 11 previously unpublished *Cryptosporidium* genome sequences (Table S1). In addition, we characterise the global distribution of *Cryptosporidium* species and subtypes, summarising the data of 743 peer-reviewed publications of cases in a total of 126 countries that used the gp60 locus for species identification and subtyping. We describe the evolutionary genomic changes of this pathogen during its association with its human host and host-range specialisation, and we estimate divergence times for the primary anthroponotic lineages. Our analyses provide a revised evolutionary scenario supporting the more recent emergence of a previously cryptic, phylogenetically-distinct anthroponotic *Cryptosporidium parvum anthroponosum* sub-species.

Results

A phylogenetic analysis of 61 neutrally-evolving coding loci across 21 *Cryptosporidium* isolates reveals the evolutionary history of human-infective taxa and identifies two discrete *C. parvum* lineages with distinct host associations, namely *C. p. parvum* (zoonotic) and *C. p. anthroponosum* (anthroponotic) (Fig. 1a; Fig. S1)[13]. Primary human-infective isolates[15] *C. hominis* and *C. parvum* form a distinct superclade with zoonotic *C. cuniculus*, a recently-identified cause of human outbreaks[16,17]. This superclade is genetically distinct from other zoonotic human-infectious *Cryptosporidium* species (*C. meleagridis*[18], *C. viatorum*[19], *C. ubiquitum*[20], *C. baileyi*[21] and *C. muris*[22]; Fig. 1a; Fig. S2; absolute divergence (d_{xy}) = 0.083 – 0.478). Within the superclade, limited genetic divergence between *C. hominis* and *C. parvum* (d_{xy} = 0.031) illustrates the recent origins of these taxa. Finally, the concatenated phylogeny provides a preliminary genotypic association between phenotypically-diverse *C. parvum* strains. Based on the host ranges of a total of 1331 isolates, *C. p. anthroponosum* UKP15 (subtype IIc-a) is almost exclusively found in humans (92.2%), whereas *C. p. parvum* UKP6 and UKP8 (subtypes IIa and IId, respectively) are more often found in ruminants than in humans (Fig. 1S). These zoonotic subtypes (UKP6 and UKP8) split off into a unique sister group (*C. p. parvum*) within the *C. parvum* clade, distinct from the anthroponotic subtype (*C. p. anthroponosum*). This switch in host association is associated with surprisingly low levels of genetic divergence (d_{xy} = 0.002), suggesting it happened recently.

Next, we undertook a meta-analysis to establish the distribution and population genetics of these *Cryptosporidium* species and subtypes based on gp60 genotyping, summarising the data of 743 peer-reviewed publications of cases in a total of 126 countries worldwide published between 2000 and 2017. The anthroponotic species *C. hominis* and *C. p. anthroponosum* are relatively more prevalent in resource poor countries (Fig. 1b,c). In contrast, the zoonotic *C. p. parvum* dominates in North America, Europe, parts of the Middle East and Australia. Even though *C. p. anthroponosum* is less prevalent in Europe (17%; 22 out of 128 cases), the mean nucleotide diversity at gp60 is significantly higher than that of *C. p. parvum* (π = 0.02954 vs. 0.00327, respectively) (Mann-Whitney test: W= 430412; $p < 10^{-5}$) (Fig. 1d). The population genetic structure differs significantly between *C. p. anthroponosum* and *C. p. parvum* (GLM:

$F_{1,79} = 47.34$, $p < 0.0001$), with *C. p. parvum* showing a strong isolation-by-distance signal, whereas there is no geographic population genetic structure for *C. p. anthroponosum* (Fig. 1e; Tables S2, S3). In Europe, *C. p. parvum* forms a geographically-structured population which shows significant isolation-by-distance (Fig. 1f,g). This suggests that gene flow within Europe shapes the genetic differentiation (F_{st}) of *C. p. parvum*, and that this pathogen is transmitted between European countries. In contrast, the high nucleotide diversity and lack of geographic structuring implies that *C. p. anthroponosum* may be introduced in Europe from genetically diverged source populations. The population genetic structure of both species is also different when analysed across a global-scale, with network analysis revealing significant sub-structuring of global populations of *C. p. parvum*, but not of *C. p. anthroponosum* (Fig. 1g,h).

Nucleotide divergence between *C. p. parvum* and *C. p. anthroponosum* is driven partly by positive selection, as evidenced by the relatively elevated ratio of Ka/Ks (> 1.0) for 44 loci (Fig. 2a; Table S4). The Ka/Ks ratio between the *C. parvum* subspecies is comparable to the Ka/Ks ratio of *C. p. parvum* and *C. hominis* comparison, and significantly higher than the Ka/Ks ratio of comparisons between other *C. p. parvum* subtypes (Fig. 2b). The signature of adaptive evolution is most apparent in the peri-telomeric genes (Fig. S4). Furthermore, frameshift-causing indels also underpin protein divergence in 130 (55.6%) and 24 (53.3%) variable *C. hominis* and *C. p. anthroponosum* amino acid coding sequences, respectively (Table S5, S6). When accounting for the size of the different chromosomal regions, indels are significantly more common in the peri-telomeric and subtelomeric regions than elsewhere in the genome (Chi-sq. test: $X^2 = 257.71$, df = 2, $p = 1.09 \times 10^{-56}$) (Fig. 2c). Genes encoding for extracellular proteins show a significantly stronger signal of positive selection than genes with a cytoplasmic protein localization (Mann-Whitney test: W = 842985, p = 0.0182) (Fig. 2d; S5), consistent with adaptations/specialisation to the human host.

Besides nucleotide substitutions and indels, genetic introgression also appears to play a prominent role in the adaptive evolution of *Cryptosporidium*. To investigate genome-wide patterns of divergence between *Cryptosporidium* lineages we aligned reads from 16 isolates to the *C. parvum* Iowa reference genome[9]. Principle component analysis based on a set high quality SNPs supports the sub-species assignments of zoonotic *C. p. parvum* and anthroponotic *C. p. anthroponosum* (Fig. 3a). Surprisingly, one sample (UKP16), identified as *C. p. parvum* based on phylogenetic analysis of 61 single copy conserved genes (Fig. 1a), appears to be highly differentiated based on genome wide SNPs (Fig. 3a). To further investigate the evolutionary history of this sample we generated phylogenetic trees in 50 SNP windows across the genome. The consensus topology of these genomic windows is shown as a "cloudogram" (Fig. 3b), which matches the concatenated analysis of conserved protein coding genes (Fig. 1a), with UKP16 most closely related to *C. p. parvum* isolates. However, many alternative topologies are also observed, indicating potential recombination between lineages (Fig. 3b). We used topology weighting[23] to visualise the distribution of topologies across the genome, focusing on evolutionary relationships between UKP16, *C. p. parvum* isolates and *C. p. anthroponosum* isolates (Fig. 3c). This analysis revealed a large region in chromosome 8 (~500 - 650Kb) where UKP16 has a sister relationship to *C. p. parvum* isolates and *C. p. anthroponosum* isolates (topo1; Fig. 3c and d). Intriguingly, this appears to be due introgression into the UKP16 genome from a highly divergent, and as yet unsampled, lineage. We draw this conclusion because the absolute divergence (d_{xy}) between UKP16 and both *C. p. anthroponosum* and *C. p. parvum* is elevated in this region, whereas divergence between *C. p. anthroponosum* and *C. p. parvum* is similar to the rest of the chromosome (Fig. 3e).

Next, we conducted a detailed analysis of genetic introgression, studying two *C. parvum parvum* isolates (UKP6 and UKP16), one *C. parvum anthroponosum* isolate (UKP15), and one *C. hominis* isolate (UKH1). A total of 104 unique recombination events are detected across these four whole genome sequences (Fig 4a; Table S7). Many recombination events involve an unknown parental sequence (i.e. donor), which is consistent with our findings for the UKP16 sample, where we identified an introgressed genomic segment from a diverged lineage (see above). These results highlight that genetic exchange is widespread across *Cryptosporidium* species. The distribution of recombination events varies markedly across chromosomes, with a disproportionately higher number of individual events detected in chromosome 6 (25.9% of total events), and a disproportionately lower number of events in chromosomes 3, 5, and 7 (Fig. S6). Another consequence of introgression is that the coalescence time between different subtypes can vary markedly within and across chromosomes, ranging from an estimated 776 to 146,415 generations ago (Table S7). Furthermore, many recombination events are detected in the peri-telomeric genes (Fig. 4a). Interestingly, of the 44 genes that appear to be under positive selection (Ka/Ks>1; see Fig. 2a), no less than 17 (38.64%) are affected by recombination. This is significantly higher than the 6.57% of genes (237 out of 3607 genes) affected by recombination that are neutrally evolving or under purifying selection (Ka/Ks<1) (Chi-square test: $X^2 = 54.51$, df = 1, p = 1.55×10^{-13}). In addition, a significantly greater number of recombination events is observed in *C. p. anthroponosum* (n=39) than in *C. hominis* (n=7) (binomial test: p = 3.12×10^{-7}) and *C. p. parvum* (n=17) (binomial test: p = 0.011) (Table S7). These analyses suggest that the genetic exchange between diverged lineages is unlikely to be a neutral process and may be fuelling adaptation in anthroponotic lineages of *Cryptosporidium*.

Finally, we estimate the divergence dates to provide the first chronological description for genetic introgression between human-infective *Cryptosporidium* spp. (Fig. 4b). The majority of introgression events between *C. p. parvum* and *C. p. anthroponosum* strains are estimated to have taken place at approximately 10-15 thousand generations ago (TGA). Only circa 6.8% of all genetic exchanges are introgression events into the *C. hominis* genome, and as expected, these events are more ancient (i.e. ~75-150 TGA). To translate generation time into years and estimate the age of the introgression events, we assume a generation time of between 48 and 96 hours[24,25], and a steady rate of transmission within host populations. The following estimates should be considered minimum estimates of divergence times because *Cryptosporidium* may be dormant outside the host. We estimate that the zoonotic *C. p. parvum* and the anthroponotic *C. p. anthroponosum* strains are likely to have recombined between 55-164 years ago, whereas we estimate that introgression events between *C. hominis* and *C. parvum* occurred between 410-1096 years ago (Fig. 4b). We show that despite genetic adaptation to specific hosts, diverged *Cryptosporidium* (sub)species continue to exchange genetic information through hybridisation within the last millennium, and that such exchange does not appear to be selectively neutral.

Discussion

Cryptosporidium is an apicomplexan parasite that can cause debilitating gastrointestinal illness in animals and humans worldwide. In order to better understand the biology of this parasite, we conducted an analysis to describe the population structuring based on 467 sequences of a highly-polymorphic locus (gp60), and we study the evolution of this parasite using 16 whole genome sequences. We demonstrate here that *C. parvum* consists of two subspecies with distinct host associations, namely *C. p. parvum* (zoonotic) and *C. p. anthroponosum* (anthroponotic) that have diverged recently. Nevertheless, the population

genetic structure differs significantly between both subspecies, with *C. p. parvum* showing a strong isolation-by-distance signal, whilst there is no clear geographic structure for *C. p. anthroponosum*. Besides the apparent differences in drift and gene flow, the divergence of both subspecies is also driven by positive selection, and the signature of adaptive evolution is comparable to that of *C. p. parvum* and *C. hominis*. Perhaps most remarkably, hybridisation has frequently led to the genetic introgression between these (sub)species. Given that such exchanges appear to be associated in particular to genes under positive selection, we believe that hybridisation plays an important role throughout the evolution of these parasites. Next, we describe *Cryptosporidium* biology with the aim to interpret and explain the population genetic and evolutionary genetic findings, placing them into the context of recent whole genome studies of other pathogens.

Our population genetic analysis detected remarkable differences between *C. p. anthroponosum* and *C. p. parvum*, both in their population genetic structure, as well as their levels of nucleotide diversity. *C. p. parvum* can cause neonatal enteritis (scour) predominantly in pre-weaned calves[26]. Given that such calves are able to produce circa 100,000 oocysts per gram of faeces, they are thought to be the primary source of subsequent infections[27]. Movement of such young animals has therefore been highly restricted by the European Union[28,29]. Adult cattle tend to be asymptomatic and shed fewer oocysts, and consequently, they are believed to be minor transmission vectors. Furthermore, long distance translocation of cattle is rare compared to human migration; just 42,515 cattle were exported to the EU from the UK[30] whereas 70.8 million overseas visits were made by UK residents in 2016[31]. Consequently, in cattle *C. p. parvum* mediated scour is unlikely to be spread by long distance migration via the livestock trade in Europe. In contrast, a significant component of human cryptosporidiosis is traveller's diarrhoea – and even where contracted domestically, the source of infection is frequently distant[32,33,34]. We propose that the difference in migration patterns between the primary hosts can explain why we find no evidence of isolation-by-distance for *C. p. anthroponosum* in Europe, whilst there is strong geographic structuring in *C. p. parvum.* Differences in the rate of gene flow can also explain the notable distinction in the nucleotide diversity between these subspecies, which is almost an order of magnitude higher in *C. p. anthroponosum* than in *C. p. parvum*. Interestingly, parasite species from the *Plasmodium* genus show the opposite pattern in that the human-infective parasite species (*P. falciparum* and *P. malariae*) have a significantly lower nucleotide diversity compared to related zoonotic malarias (*P. reichenowi* and *P. malariae*-like)[35,36]. In this example, the lack of diversity in human-infective species has been interpreted as evidence for their recent population expansions. In *C. p. anthroponosum*, however, our population genetic analysis suggests that nucleotide diversity in the European population has been restored by introduction of novel genetic variation through immigration from diverged source populations outside Europe, as well as by genetic introgression.

Besides gene flow, our analysis identifies a strong signal of hybridisation between diverged strains or species, and we suggest that such genetic exchange between diverged taxa (i.e. genetic introgression) may also have contributed to the rapid restoration of diversity of *C. p. anthroponosum.* We detect 104 unique recombination events and estimate that the genetic exchanges have taken place relatively recently, i.e. within the last millennium or ~100,000 generations. This implies that hybridisation plays an important role in the biology of *Cryptosporidium*, and that this complex of *Cryptosporidium* species is coevolving in the presence of recent or continued genetic exchange. This interpretation is consistent with the growing body of evidence suggesting that hybridisation of diverged strains plays an important role in pathogen evolution[6,37]. Hybridisation can lead to the sharing of conserved

haploblocks across distinct phylogenetic lineages or (sub)species. Such mosaic-like genomes have been observed also in other human pathogens like *Toxoplasma gondii*[7], as well some plant pathogens such as the oomycete, *Albugo candida*[38]. Hybridisation can only occur, however, when different strains are in physical contact with one another. Unlike *A. candida*, which appears to suppress the host's immune response and facilitate coinfections[38], challenge experiments with human-infective isolates have shown that different *Cryptosporidium* species compete with each other within the host. For example, the *C. parvum parvum* strain GCH1 (subtype IIa) was shown to rapidly outcompete *C. hominis* strain TU502 (subtype Ia) during mixed infections in piglets[39]. Nevertheless, mixed species infections or intra-species diversity in *Cryptosporidium* have been identified in a large number (n = 55) of epidemiological surveys of cryptosporidiosis conducted in the period between 2005 – 2015[40]. As with *A. candida*, during the potentially brief periods of coinfections, hybridisation between distinct *Cryptosporidium* lineages may take place within a single host. In turn, this could facilitate the genetic exchange between the diverged lineages and contribute to the (virulence) evolution of *Cryptosporidium*. Introgression from an unidentified source into chromosome 8 of isolate UKP16 illustrates the diversity of the genepool that is able to exchange genetic variation, and it highlights the need for whole genome sequence studies for our understanding of *Cryptosporidium* biology. Interestingly, the distribution of recombination events varies markedly across chromosomes, a pattern observed also in other pathogens such as *T. gondii*[7]. Most remarkably, however, we found that in *Cryptosporidium* genes with a signature of positive selection were significantly more likely to be located in recombination blocks than neutrally evolving genes and genes under purifying selection. Our analyses thus suggest that such exchange is unlikely to be a neutral process, and that the recent emergence of the specialised anthroponotic subspecies such as *C. p. anthroponosum* might be fuelled by relatively recent, and possibly ongoing, "adaptive introgression"[37]. We estimate that these founding introgression events in the divergence of zoonotic *C. p. parvum* from the anthroponotic *C. p. anthroponosum* began 55-164 years ago, whereas those between *C. hominis* and *C. parvum* occurred between 410-1096 years ago timing which is consistent with reduced livestock contact and increased human population densities – conditions providing a continued selection pressure for the emergence of new human adapted pathogens from zoonotic origins.

Methods

Systematic Review

A human cryptosporidiosis prevalence database was constructed using data from peer-reviewed publications retrieved using the search term "Cryptosporidium" from PubMed (https://www.ncbi.nlm.nih.gov/pubmed) published between 2000-2017. After filtering (see SI Methods), the final dataset consisted of 743 publications of human *Cryptosporidium* infections in 126 countries.

Empirical Data

Whole genome sequence (WGS) data for *C. hominis* UKH1 and *C. meleagridis* UKMEL 1 were retrieved from the *Cryptosporidium* genetics database resource CryptoDB (www.cryptoDB.org)[41]. The remaining 19 *Cryptosporidium spp*. WGS datasets were obtained from clinical isolates[8] (see Table S1).

Concatenated Phylogenetic Analysis

61 neutrally-evolving loci (Ka/Ks = 0.2-0.6; 93.0-98.0% nucleotide IDs) between *C. parvum* UKP6 and *C. hominis* UKH4 were concatenated. A concatenated approach targeting neutral loci was used in lieu of the well-known gp60 subtyping locus, as this highly recombinant locus frequently produces phylogenies that do not correlate with genome-wide divergence (Fig. S7)[42]. Orthologous protein coding sequences from the human-infective WGS UKP6 and UKH4 were extracted (Table S10), and aligned using ClustalW. The Maximum Likelihood phylogeny was constructed with the Dayhoff substitution model, the Nearest-Neighbour-Interchange method and 2,000 bootstraps[43]. Divergence statistics between lineages were calculated using MEGA7[43].

Whole Genome Comparisons

Parallel whole genome comparative analyses were performed between a zoonotic *C. p. parvum* IIaA15G2R1-subtype WGS (UKP6), anthroponotic *C. p. anthroponosum* IIcA5G3a-subtype (UKP15), and anthroponotic *C. hominis* IaA14R3-subtype (UKH4). CDS nucleotide divergence was evaluated by cross-blasting CDS datasets locally (BLOSUM62 substitution matrix; BioEdit)[44]. Amino acid identities and indels resulting in frameshift were identified using EMBOSS Stretcher[45]. Selection was identified by calculating Ka/Ks in CodeML of PAML[46], and NaturalSelection.jl (https://github.com/BioJulia/NaturalSelection.jl). Sliding window Ka/Ks analyses, indel characterisations, and F_{ST} calculations were performed in DnaSP 5.10.1[47]. Putative protein function was evaluated using the UniProt BLASTp function (cut-off E-value <10e-5)[48], and putative protein localization was estimated using WoLF PSORT[49].

Phylogenomic analysis

Sequence reads of 21 *Cryptosporidium* isolates (Table S1) were aligned to the *C. parvum* Iowa[9] reference genome and SNPs identified (see SI Methods). Pseudoreferences were generated with filtered biallelic SNPs inserted using GATK FastaAlternateReferenceMaker[50]. Principle component analysis of *C. p. parvum* and *C. p. anthroponosum* isolates was performed with SNPrelate[51]. Population genetic statistics the fixation index (F_{ST}), absolute divergence (d_{xy}) and nucleotide diversity (π) were estimated in 50 Kb sliding windows (10 Kb step size) across the genome. Maximum likelihood phylogenies were estimated for 50 SNP windows across the genome using RAxML[52]. Topology weighting[23] was used to investigate the distribution of phylogenetic relationships across the genome with each isolate assigned to one of four groups (*C. p. parvum*, *C. p. anthroponosum*, UKP16 and outgroup samples (*C. hominis* and *C. cuniculus*). Ultrametric phylogenetic trees were made using the *chronopl* function in APE[53], and a consensus phylogeny was generated.

Recombination Analysis

Recombination signals due to introgression were detected using RDP4[54]. Automated detection algorithms RDP, GENECONV, Bootscan, Maxchi, and Chimaera were run with default values. Alternative call (AC) values of all bases in the four isolates that were studied in the genetic introgression analysis (UKH1, UKP6, UKP15 and UKP16) to validate that they comprised single subtype infections (Fig. S8).

Dating introgression events

Hybridization dating was estimated for introgressed regions in HybridCheck[55]. The HKY85 substitution model with a SNP mutation rate of $\mu=10^{-8}$ per generation was assumed, based on the observed nucleotide divergence between two outbreak WGS sampled seven days apart (Table S8). To convert generations into time, we assumed a factor of 12 autoinfective offspring per parental oocyst *in vivo* (Fig. S9). Furthermore, past infectivity studies revealed a population expansion of 3-5 new generations, and an estimated life cycle duration of 48-96h per infection (Table S9)[60,61]. This estimate is longer than previous estimates (12-14h)[56], but consistent with estimates of 72h from a cell culture experiment[57]. The reported estimates of time may be underestimated if oocysts remain dormant in the environment between infections of different host individuals.

Population Genetic Analysis

A total of 467 gp60 sequences collected in 43 countries were used to analyse the population structure of *C. p. parvum* UKP6 (N=361) and *C. p. anthroponosum* UKP15 (N=106) (see SI Methods). Population genetic structure was visualised using Fluxus network using median joining setting[58]. Isolation-by-distance analysis was performed using a regression analysis of the genetic distance (Kxy) between isolates and geographic distance between the sampling locations. Differences between chromosomes, chromosomal regions, recombinant regions and genes in the number of SNPs, indels, and recombination events were tested with Chi-square and binomial tests. Differences in nucleotide substitution patterns, indels and recombination events between taxa were analysed using Mann-Whitney test and ANOVAs. All tests were conducted in R (R Core Team)[59] and Minitab 12.1.

Data availability

All WGS data used in this paper is available publically and for free via the NCBI server (https://www.ncbi.nlm.nih.gov/) or CryptoDB (http://cryptodb.org/cryptodb/). The accession codes for the data are provided in Table S1.

Author's contributions

KT, RC, PH, JN and CvO conceived the study. JN and CvO designed the analyses. JN, JP, GR, MS, PH, KT and RC were involved in the acquisition of data. JN conducted the meta-analysis. JN and CvO conducted the evolutionary genetic analyses with input of TM for the phylogenetic and BW for the recombinant analyses. JN and CvO drafted the submitted manuscript. All authors contributed to revising the draft, had full access to all the data and read and approved the final manuscript.

Acknowledgements

This work was supported with funds awarded to KT and RMC from the FP7 KBBE EU project AQUAVALENS, grant agreement 311846 from the European Union awarded to PH, and a Biotechnology and Biological Sciences Research Council (BBSRC) (BB/N02317X/1) awarded to CvO, as well as support by the Earth & Life Systems Alliance (ELSA). P.R.H. is supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections at the University of Liverpool, in partnership with Public Health England (PHE), and in collaboration with University of East Anglia, University of Oxford, and the Institute of Food Research. Professor Hunter is based at the University of East Anglia. The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research, the Department of Health, or Public Health England. We thank Gregorio Pérez-Cordóna for VNTR validation of isolates, and we thank the three reviewers for their helpful comments.

Competing Interests

The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1]Liu, L. *et al*. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet* **379**, 2151-2161 (2012).

[2]Kotloff, K. L. *et al*. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet* **382**, 209-222 (2013).

[3]Widmer, G., & Sullivan, S. Genomics and population biology of *Cryptosporidium* species. *Parasite Immunol.* **34**, 61-71 (2012).

[4]Mazurie, A. *et al*. Comparative genomics of *Cryptosporidium*. *Int. J. Genomics* **2013**, 832756 (2013).

[5]Bushell, E. *et al*. Functional profiling of a *Plasmodium* genome reveals an abundance of essential genes. *Cell* **170**, 260-72 (2017).

[6]McMullan, M. *et al*. Evidence for suppression of immunity as a driver for genomic introgressions and host range expansion in races of *Albugo candida*, a generalist parasite. *eLife* **4** (2015).

[7]Lorenzi, H. *et al*. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic *Toxoplasma gondii* genomes. *Nature Commun.* **7**, 10147 (2016).

[8]Hadfield, S. J. *et al*. Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples. *BMC Genomics* **16**, 1-12 (2015).

[9]Abrahamsen, M. S. *et al*. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* **304**, 441-445 (2004).

[10]Xu, P. *et al*. The genome of *Cryptosporidium hominis*. *Nature* **431**, 1107-1112 (2004).

[11]Bouzid, M., Hunter, P. R., Chalmers, R. M., & Tyler, K. M. *Cryptosporidium* pathogenicity and virulence. *Clin. Microbiol. Rev.* **26**, 115-134 (2013).

[12]Widmer, G. *et al*. Comparative genome analysis of two *Cryptosporidium parvum* isolates with different host range. *Infect. Genet. Evol.* **12**, 1213-1221 (2012).

[13]Guo, Y. *et al*. Comparative genomic analysis reveals occurrence of genetic recombination in virulent *Cryptosporidium hominis* subtypes and telomeric gene duplications in *Cryptosporidium parvum*. *BMC Genomics* **16**, 1-18 (2015).

[14]Li, N. *et al*. Genetic recombination and *Cryptosporidium hominis* virulent subtype IbA10G2. *Emerg. Infect. Dis.* **19**, 1573-82 (2013).

[15]Xiao, L. & Ryan U. M. Cryptosporidiosis: an update in molecular epidemiology. *Curr. Opin. Infect. Dis.* **17**, 483-90 (2004).

[16]Puleston, R. L. *et al*. The first recorded outbreak of cryptosporidiosis due to *Cryptosporidium cuniculus* (formerly rabbit genotype), following a water quality incident. *J. Water Health* **12**, 41-50 (2014).

[17]Koehler, A. V., Whipp, M. J., Haydon, S. R. & Gasser, R. B. *Cryptosporidium cuniculus* - new records in human and kangaroo in Australia. *Parasit. Vectors* **7**, 492 (2014).

[18]Wang, Y. *et al*. Population genetics of *Cryptosporidium meleagridis* in humans and birds: evidence for cross-species transmission. *Int. J. Parasitol*. **44**, 515-21 (2014).

[19]Koehler, A. V. *et al*. *Cryptosporidium viatorum* from the native Australian swamp rat Rattus lutreolus - An emerging zoonotic pathogen? *Int. J. Parasitol. Parasites Wildl*. **7**, 18-26 (2018).

[20]Li, N. *et al*. Subtyping *Cryptosporidium ubiquitum*, a zoonotic pathogen emerging in humans. *Emerg. Infect. Dis.* **20**, 217-24 (2014).

[21]Joachim, A. Human cryptosporidiosis: an update with special emphasis on the situation in Europe. *J. Vet. Med. B Infect. Dis. Vet. Public Health* **51**, 251-9. (2004).

[22]Chappell, C. L. et al. *Cryptosporidium muris*: infectivity and illness in healthy adult volunteers. *Am. J. Trop. Med. Hyg*. **92**, 50-5 (2015).

[23]Martin, S. H. & Van Belleghem, S. M. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* **206**, 429-438 (2017).

[24]Okhuysen, P. C. *et al*. Infectivity of a *Cryptosporidium parvum* isolate of cervine origin for healthy adults and interferon-gamma knockout mice. *J. Infect. Dis*. **185**, 1320-5 (2002).

[25]Chappell, C. L. *et al*. *Cryptosporidium meleagridis*: infectivity in healthy adult volunteers. *Am. J. Trop. Med. Hyg*. **85**, 238-42 (2011).

[26]Santín, M., Trout, J. M., & Fayer, R. A longitudinal study of cryptosporidiosis in dairy cattle from birth to 2 years of age. *Vet. Parasitol.* **155**, 15-23 (2008).

[27]Current, W. L. Cryptosporidiosis. *J. Am. Vet. Med. Assoc*. **187**, 1334-8 (1985).

[28]Animal Transport Guides, Transport of calves. (2017). at: <http://animaltransportguides.eu/>.

[29]Defra,. PB 12544a: Welfare of Animals During Transport. (2011).

[30]LAres, E., & Ward, M. Live animal exports. *Commons Library Briefing.* **8031** (2017).

[31]ONS. Travel Trends: 2016. (2017). <https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/articles/traveltrends/2016>.

[32]Jelinek, T. *et al*. Prevalence of infection with *Cryptosporidium parvum* and *Cyclospora cayetanensis* among international travellers. *Gut* **41**, 801-804 (1997).

[33]Nair, P. *et al*. Epidemiology of cryptosporidiosis in North American travelers to Mexico. *Am. J. Trop. Med. Hyg.* **79**, 210-4 (2008).

[34]Chalmers, R. M. *et al*. Geographic linkage and variation in *Cryptosporidium hominis*. *Emerg. Infect. Dis.* **14**, 496-8 (2008).

[35]Sundararaman, S. A. *et al*. Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. *Nat. Commun*. **22**, 11078 (2016).

[36]Rutledge, G. G. *et al*. *Plasmodium malariae* and *P. ovale* genomes provide insights into malaria parasite evolution. *Nature* **542**, 101-104 (2017).

[37]King, K. C., Stelkens, R. B., Webster, J. P., Smith, D. F. & Brockhurst, M. A. Hybridization in parasites: consequences for adaptive evolution, pathogenesis, and public health in a changing world. *PLoS Pathog.* **11** (2015).

[38]Jouet, A. *et al*. *Albugo candida* race diversity, ploidy and host□associated microbes revealed using DNA sequence capture on diseased plants in the field. *New Phytol.* doi: 10.1111/nph.15417 (2018).

[39]Akiyoshi, D. E., Mor, S. & Tzipori, S. Rapid displacement of *Cryptosporidium parvum* type 1 by type 2 in mixed infections in piglets. *Infect. Immun.* **71**, 5765-71 (2003).

[40]Grinberg, A. & Widmer, G. *Cryptosporidium* within-host genetic diversity: systematic bibliographical search and narrative overview. *Int. J. Parasitol*. **46**, 465-71 (2016).

[41]Puiu, D. *et al*. CryptoDB: the *Cryptosporidium* genome resource. *Nucleic Acids Res.* **32**, D329-31 (2004).

[42]Feng, Y., Ryan, U. M. & Xiao, L. Genetic diversity and population structure of *Cryptosporidium*. *Trends Parasitol.* **34**, 997-1011 (2018).

[43]Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol*. **33**, 1870-4 (2016).

[44]Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95-98 (1999).

[45]Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. **16**, 276-7 (2000).

[46]Suyama, M., Torrents, D. & Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. **34**, W609-12 (2006).

[47]Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-2 (2009).

[48]Apweiler, R. *et al*. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*. **32**, D115-9 (2004).

[49]Horton, P. *et al*. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* **35**, W585-7 (2007).

[50]DePristo, M. A. *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491-8 (2011).

548 [51]Zheng, X. *et al*. A High-performance Computing Toolset for Relatedness and Principal
549 Component Analysis of SNP Data. *Bioinformatics* **28**, 3326-3328 (2012).

550 [52]Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
551 large phylogenies. *Bioinformatics* **30**, 1312-3 (2014).

552 [53]Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R
553 language. *Bioinformatics* **20**, 289-90 (2004).

554 [54]Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and
555 analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, vev003 (2015).

556 [55]Ward, B. J. & van Oosterhout, C. HYBRIDCHECK: software for the rapid detection,
557 visualization and dating of recombinant regions in genome sequence data. *Mol. Ecol. Resour*.
558 **16**, 534-9 (2016).

559 [56]Fleming, R. *Cryptosporidium*: Could It Be in Your Water? Ontario Ministry of Agriculture,
560 Food, and Rural Affairs (2015).

561 [57]Current, W. L. & Haynes, T. B. Complete development of *Cryptosporidium* in cell culture.
562 *Science* **224**, 604-5 (1984).

563 [58]Bandelt, H. J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific
564 phylogenies. *Mol. Biol. Evol*. **16**, 37–48 (1999).

565 [59]R Core Team. R: A language and environment for statistical computing. R Foundation for
566 Statistical Computing, Vienna, Austria (2013). URL http://www.R-project.org/.

567 [60]Kosek, M., Alcantara, C., Lima, A. A. & Guerrant, R. L. Cryptosporidiosis: an update.
568 *Lancet Infect. Dis*. **1**, 262-9 (2001).
569
570 [61]O'Hara, S. P. & Chen, X. M. The cell biology of *Cryptosporidium* infection. *Microbes*
571 *Infect*. **13**, 721-30 (2011).

572

Legends to Figures

Figure 1

a, Concatenated phylogeny of 16 human-infective *Cryptosporidium spp*. The maximum likelihood phylogeny is based on a 142,452 bp alignment of 61 loci (Table S10) and 2,000 bootstrap replications. Unique UK-identifiers show species group, specific gp60 subtype, and prevalent host type(s) (Table S1, Fig. S1). **b,c,** Relative global distribution of human cryptosporidiosis due to *C. parvum* (orange) versus *C. hominis* (blue) based on a systematic review of 743 peer-reviewed publications (Dropbox). Relative proportion of global *C. parvum* human cryptosporidiosis due to zoonotic *C. p. parvum* IIa (green) versus anthroponotic *C. p. anthroponosum* IIc-a (purple) based on a systematic review of 84 peer-reviewed publications. **d,** Nucleotide diversity (π) within European *C. p. parvum* (IIa) (green, n=96; Min=0.000000, 1st Qu.=0.001374, Median=0.002762, Mean=0.003244, 3rd Qu.=0.004169, Max=0.006970) and *C. p. anthroponosum* (IIc-a) (purple, n=22; Min=0.000000, 1st Qu.=0.002124, Median=0.043951, Mean=0.029704, 3rd Qu.=0.046250, Max=0.061045) populations. **e**, The genetic distance (Kxy) between *C. p. parvum* (n=345) isolates is strongly correlated with geographic distance (Regression $F_{1,26}$=40.63, p=0.000000944, R^2=61.0%), whilst there is no isolation-by-distance signal detected for *C. p. anthroponosum* (n=106) isolates ($F_{1,16}$=1.477, p=0.242). **f**, *C. p. parvum* (IIa) isolates show an isolation-by-distance signal, as is illustrated by the positive slope of the regression line between genetic differentiation (Fst) and geographic distance (Regression: R^2-adj.=58.3%, $F_{1,8}$=13.60, p=0.006). This signal suggests there is some gene flow within Europe. No isolation-by-distance was found for *C. p. anthroponosum* (IIc-a) in Europe. Combined with significantly higher nucleotide diversity, this suggests that *C. p. anthroponosum* infections arrive from outside Europe, rather than being transmitted within Europe. **g,h,** Fluxus network of global *C. p. parvum* (IIa) and *C. p. anthroponosum* (IIc-a) GenBank-submitted gp60 sequences show significant sub-structuring of global populations of *C. p. parvum* IIa isolates, and absence of structure between or within regional populations of *C. p. anthroponosum* IIc-a.

Figure 2

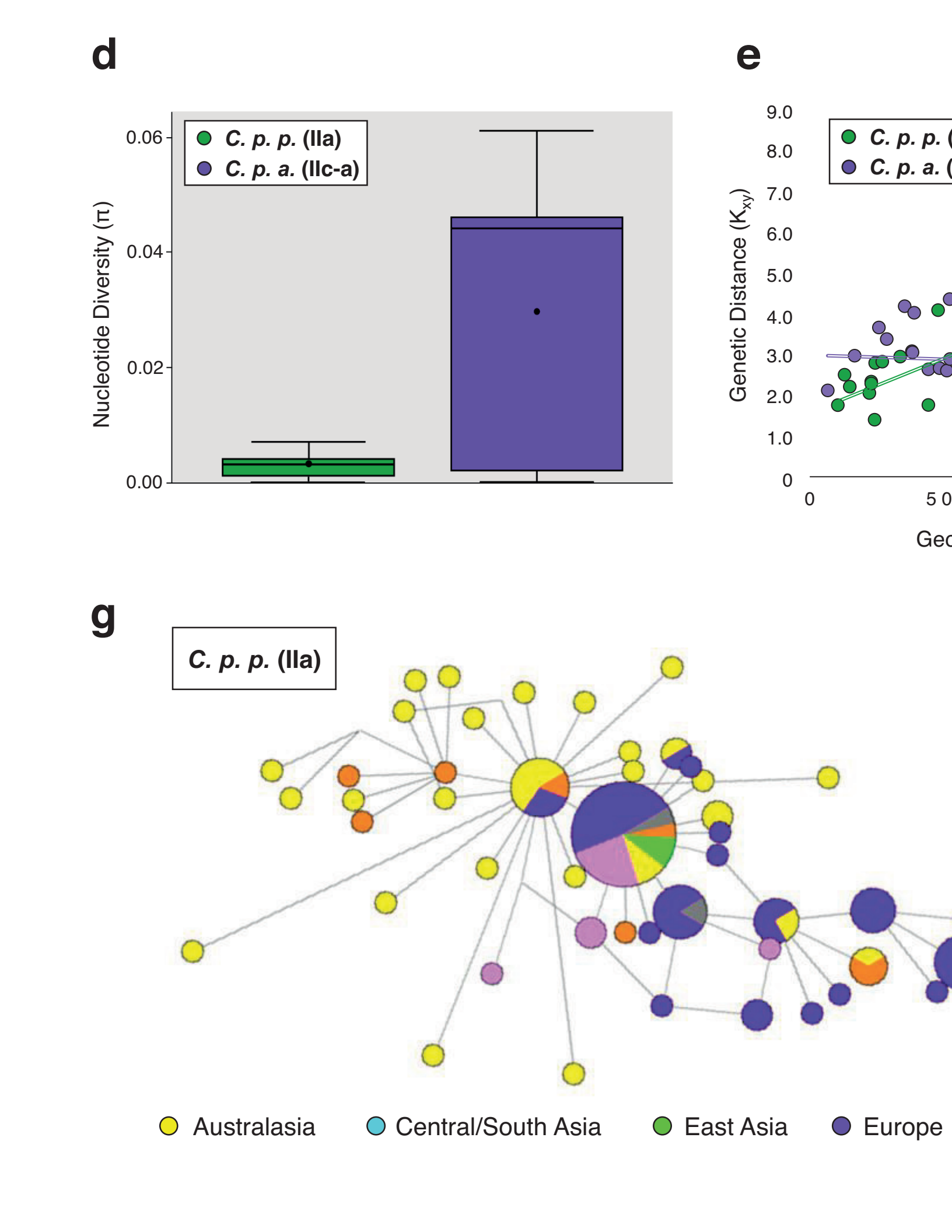
a,**b**, Selective pressures (Ka/Ks) and nucleotide distances (π) generated gene-by-gene between and within zoonotic and anthroponotic *Cryptosporidium* species groups. Zoonotic *C. p. parvum* UKP6 genomics coding sequences (CDSs) are here compared to zoonotic *C. p. parvum* UKP8 (green; Min=0.00000, 1st Qu.=0.00000, Median=0.00000, Mean=0.1613, 3rd Qu.=0.00000, Max=1.00000), anthroponotic *C. parvum parvum* UKP16 (yellow; Min=0.00000, 1st Qu.=0.00000, Median=0.00000, Mean=0.17991, 3rd Qu.=0.09046, Max=1.00000), anthroponotic *C. p. anthroponosum* UKP15 (red; Min=0.00000, 1st Qu.= 0.00000, Median=0.00000, Mean=0.2169, 3rd Qu.=0.2219, Max=1.00000), and anthroponotic *C. hominis* UKH4 (blue; Min=0.00000, 1st Qu.=0.05924, Median=0.11785, Mean=0.13858, 3rd Qu.=0.18854, Max=1.00000). Distribution of global Ka/(Ka+Ks) values for each comparison are shown, and differences were assessed statistically (One-way ANOVA, $F_{12,727} = 31.34$, P<3.567e-20, n=3465 CDSs). **c**, Sliding window analysis of triplet (brown) and non-triplet (green) insertion and deletion (indel) events between two samples, i.e. *C. parvum parvum* UKP6 and *C. parvum anthroponosum* UKP15. Composite results for 20 kb-wide sliding windows across chromosomes 1, 2, 4, 6, and 8 are shown. Peri-telomeric genes (T) and subtelomeric genes (S) have significantly more triplet and non-triplet indels than non-telomeric (NT) genes (Chi-sq. test, X^2=38.535, df=2, p=4.29x10^{-9}; X^2=226.078, df=2, p=8.09e^{-50}, respectively). **d**, Comparative selective pressure analysis between *C. p. parvum* UKP6 and *C. p. anthroponosum* UKP15 coding sequences with contrasting protein localizations. The range of Ka/(Ka+Ks) between all (n=3465; Min=0.00000, 1st Qu.=0.00000, Median=0.1416, Mean=0.3058, 3rd Qu.=0.3989, Max=1.00000) CDSs, CDSs annotated as having a cytoplasmic protein localization (n=1152; Min=0.00000, 1st Qu.=0.00000, Median=0.1110, Mean=0.2980, 3rd Qu.=0.3705, Max=1.00000), and CDSs annotated as having an extracellular localization (n=333; Min=0.00000, 1st Qu.=0.00000, Median=0.1973, Mean=0.4180, 3rd Qu.= 1.00000, Max=1.00000) are represented by a violin plot. CDSs with extracellular localisation experience significantly more positive selection than cytoplasmic CDSs, as evidenced by their higher Ka/(Ka+Ks) value (two-sided Mann-Whitney test, W=842985, p=0.0182). In addition, 17 out of 333 (5.1%) extracellular CDSs have a Ka/Ks larger than unity, compared to just 21 out of 3236 (0.6%) cytoplasmic CDSs (Chi-sq. test: X^2=53.8, d.f.=1, p=1.675e-12).

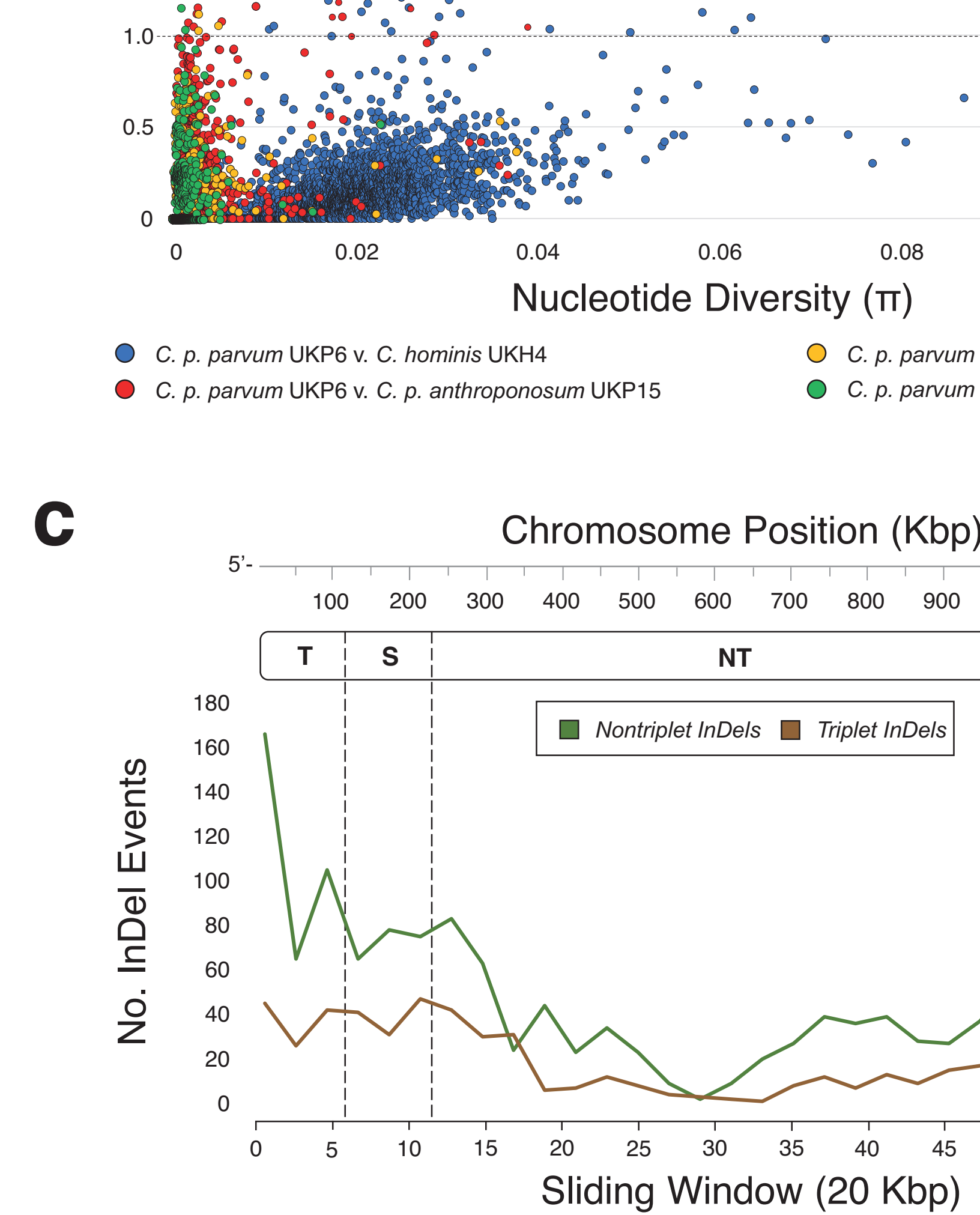
Figure 3

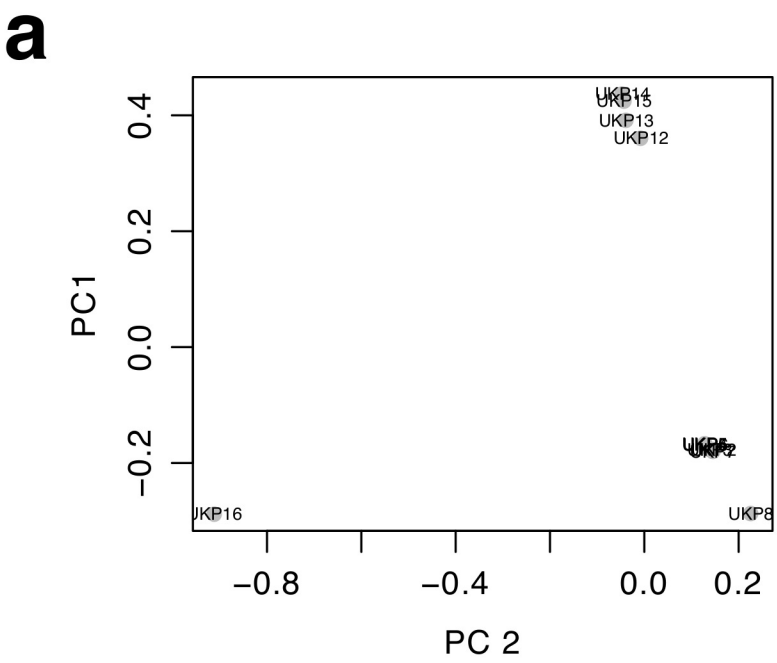
a, Principle component analysis of *C. p. parvum* and *C. p. anthroponosum* isolates based on 1,476 high quality SNPs retained after pruning based on linkage disequilibrium. **b**, A "cloudogram" of 1,324 trees showing phylogenomic relationships between WGS of anthroponotic *Cryptosporidium* isolates. Maximum likelihood trees were estimated for non-overlapping 50 SNP genomic windows across the *C. parvum* Iowa II reference genome (grey). The consensus phylogeny is shown in black. Isolates belonging to *C. p. parvum* and *C. p. anthroponosum* sub-species fall into two monophyletic groups, C. hominis/C.cuniculus isolates are included as an outgroup (OG). **c**, Topology weighting was used to explore the genome-wide distribution of phylogenetic relationships between the two *C. parvum* subspecies, a putatively introgressed isolate (UKP16) and an outgroup (*C. hominis* isolates and a single *C. cuniculus* isolate) using the 50 SNP fixed window trees. All possible topologies of the ingroup taxa are shown in the top panel, the lower panel shows the genome-wide average weighting of each topology. **d**, The distribution of topology weightings across chromosome 8 (colours as per c) reveals a putatively introgressed region between 500Kb and 650Kb. **e**, Absolute divergence (d_{xy}) between *Cryptosporidium* sub-species and the putatively introgressed isolate UKP16 in 50 Kb sliding windows (10Kb step size) across chromosome 8 of the *C. parvum* Iowa II reference genome.

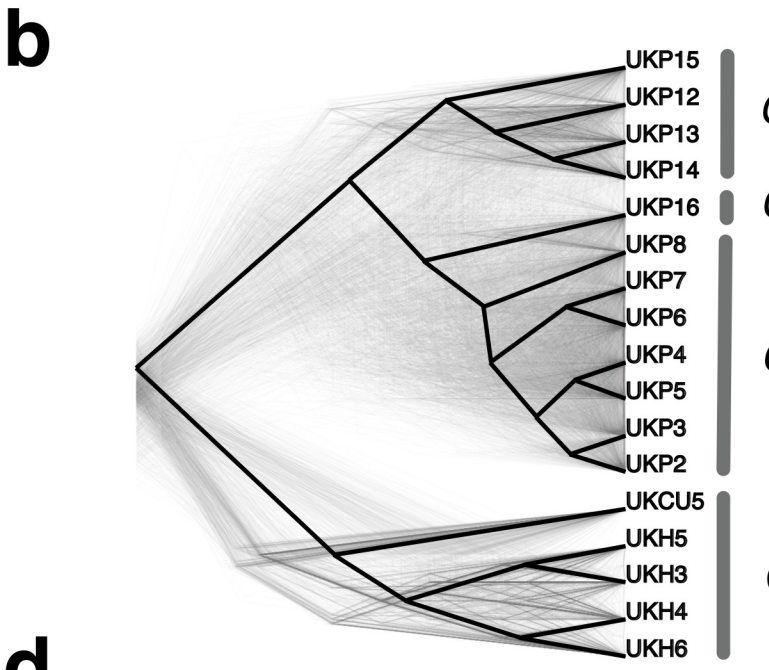
Figure 4

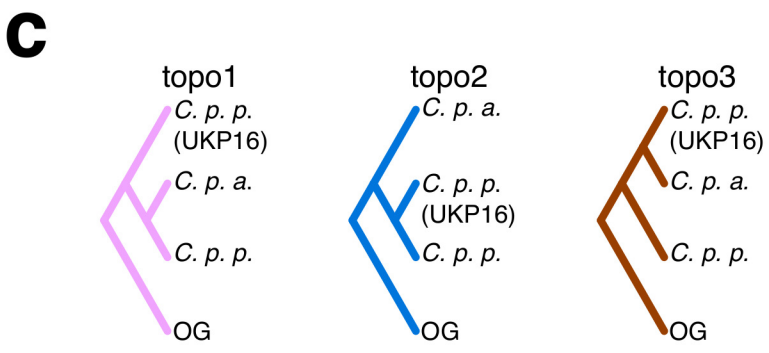
a, Genomic recombinant events in anthroponotic *Cryptosporidium spp*. WGS. Size and location of recombinant fragments detected by RDP4 are illustrated for recombination between *C. p. parvum* UKP6 and *C. p. parvum* UKP16 (yellow), *C. p. parvum* UKP6 and *C. p. anthroponosum* UKP15 (pink), *C. p. parvum* UKP16 and *C. p. anthroponosum* UKP15 (turquoise), *C. p. parvum* UKP6 and *C. hominis* UKH1 (green), *C. p. anthroponosum* UKP15 and *C. hominis* UKH1 (blue), and *C. p. parvum* UKP16 and *C. hominis* UKH1 (peach). Recombination events with unknown major or minor parentage are additionally represented (grey). Individual recombination events are detailed in Table S7. **b**, Estimated dates of introgression events between anthroponotic and zoonotic *Cryptosporidium spp*.. The range of estimated introgression times (thousands of generations ago) are given for introgression events between zoonotic *C. p. parvum* (UKP6) and anthroponotic *C. p. anthroponosum* (UKP15) – n=45, Min=7369, 1st Qu.=9218, Median=11486, 3rd Qu=13045, Max=17914 , and for introgression events between zoonotic *C. p. parvum* (UKP6) and anthroponotic *C. hominis* (UKH1) – n=33, Min=64655, 1st Qu.=77337, Median=95974, Mean=103281, 3rd Qu.117130, Max=188341. Minimum, mean, and maximum generation numbers were converted into units of time (years) for both 48- and 96-hour life cycle estimates.

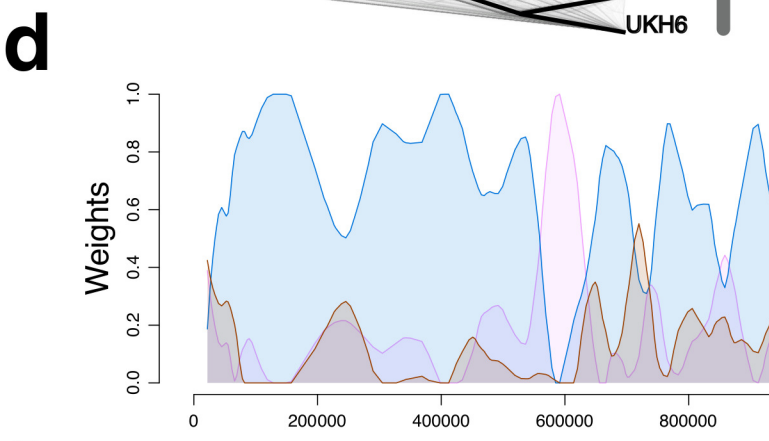

d
C. p. p. (IIa)
C. p. a. (IIc-a)
Nucleotide Diversity (π)
0.06
0.04
0.02
0.00
e
C. p. p. (
C. p. a. (
Genetic Distance (Kxy)
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0
0
5 0
Geo
g
C. p. p. (IIa)
Australasia
Central/South Asia
East Asia
Europe

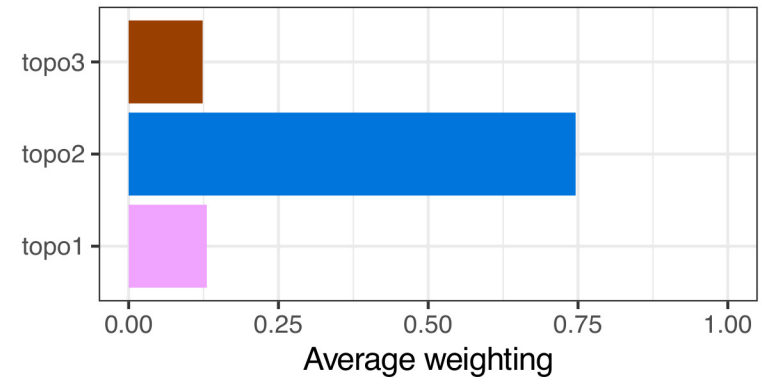

1.0
0.5
0
0
0.02
0.04
0.06
0.08
Nucleotide Diversity (π)
C. p. parvum UKP6 v. C. hominis UKH4
C. p. parvum UKP6 v. C. p. anthroponosum UKP15
C. p. parvum
C. p. parvum
c
Chromosome Position (Kbp)
5'-
100
200
300
400
500
600
700
800
900
T
S
NT
Nontriplet InDels
Triplet InDels
No. InDel Events
180
160
140
120
100
80
60
40
20
0
0
5
10
15
20
25
30
35
40
45
Sliding Window (20 Kbp)

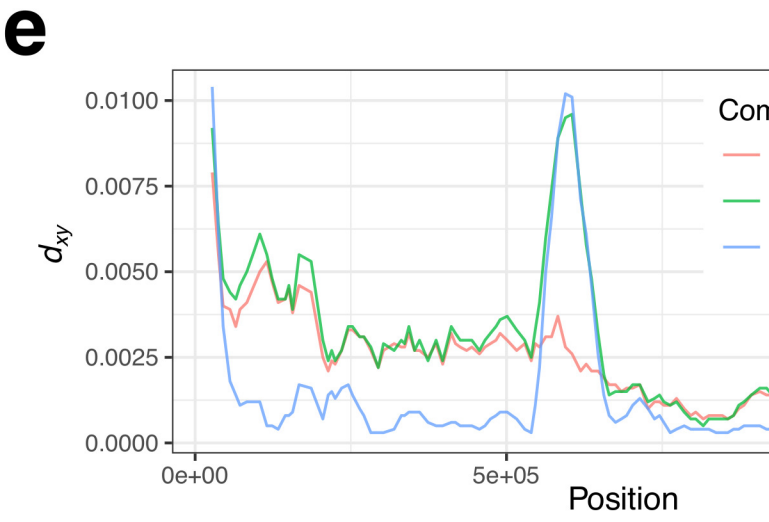

a
PC1
0.4
0.2
0.0
−0.2
UKP14
UKP15
UKP13
UKP12
UKP16
UKP8
−0.8
−0.4
0.0
0.2
PC 2


b
UKP15
UKP12
UKP13
UKP14
UKP16
UKP8
UKP7
UKP6
UKP4
UKP5
UKP3
UKP2
UKCU5
UKH5
UKH3
UKH4
UKH6


c
topo1
C. p. p.
(UKP16)
C. p. a.
C. p. p.
OG
topo2
C. p. a.
C. p. p.
(UKP16)
C. p. p.
OG
topo3
C. p. p.
(UKP16)
C. p. a.
C. p. p.
OG


d
Weights
0.0
0.2
0.4
0.6
0.8
1.0
0
200000
400000
600000
800000


topo3
topo2
topo1
0.00
0.25
0.50
0.75
1.00
Average weighting

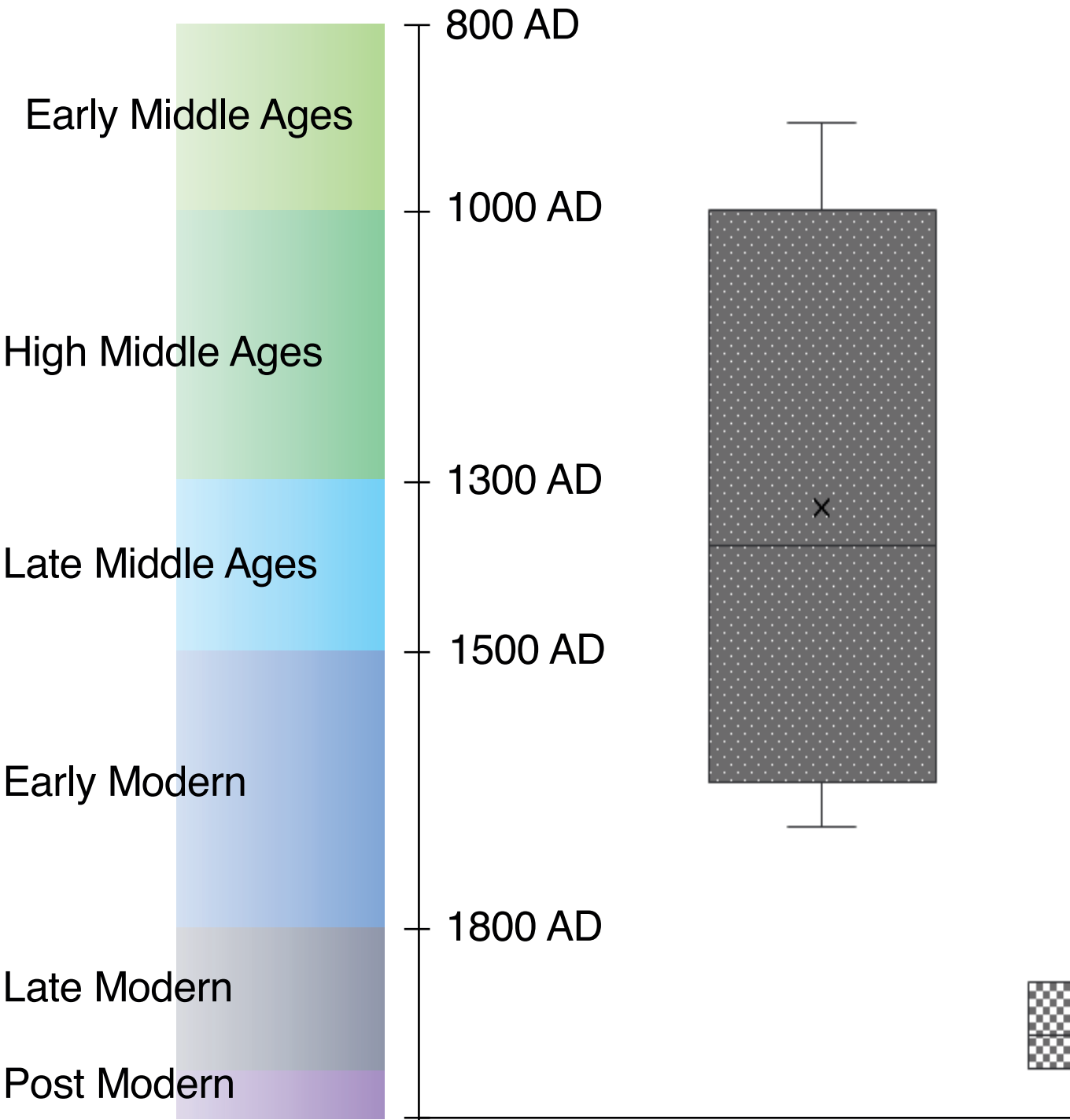

e
d_{xy}
0.0100
0.0075
0.0050
0.0025
0.0000
0e+00
5e+05
Position
Com

Po

Major Parent:Minor Parent

- *UKP6:UKP16*
- *UKP6:UKP15*
- *UKP15:UKP16*
- *UKP6:UKH1*
- *UKP15:UKH1*
- *UKP16:UKH1*
- *Minor or major parent unknown*

b



Supplementary Table 1

Summary statistics of 25 human-infective *Cryptosporidium spp*. genome projects including 23 whole genome sequences (WGS).

SPECIES	gp60 SUBTYPE	STANDARD ID	SOURCE	ACCESSION	WGS SIZE (bp)	N50 (Mb)	Host	Co
C. cuniculus	VbA37	UKCU2*	This study	PRJNA315496	9,183,765	1.806	Human	201
C. cuniculus	VaA31	UKCU5#	This study	PRJNA492839	Not Assembled		Human	201
C. hominis	IbA10G2	UKH1‡	Widmer, G.¥	CryptoDB.org	9,141,398	0	Human	201
C. hominis	IbA10G2	UKH3#	Hadfield *et al*[1]	PRJNA253834	9,136,308	0.060	Human	201
C. hominis	IaA14R3	UKH4*	Hadfield *et al*[1]	PRJNA253838	9,158,297	0.167	Human	201
C. hominis	IbA10G2	UKH5#	Hadfield *et al*[1]	PRJNA253839	9,179,731	0.168	Human	201
C. hominis	IdA30	UKH6#	This study	PRJNA492838	Not Assembled		Human	201
C. meleagridis	IIIbA22G1	UKMEL1*	Widmer, G.¥	CryptoDB.org	8,973,224	0	Human	201
C. meleagridis	IIIgA23G3	UKMEL3*	This study	PRJNA315502	8,732,077	0.062	Human	201
C. meleagridis	IIIhA7	UKMEL4*	This study	PRJNA315503	8,811,811	0.025	Human	201
C. parvum parvum	IIaA19G1R2	UKP2*	Hadfield. *et al*[1]	PRJNA253836	9,104,817	0.034	Human	201
C. parvum parvum	IIaA18G2R1	UKP3*	Hadfield *et al*[1]	PRJNA253840	9,085,662	0.126	Human	201
C. parvum parvum	IIaA15G2R1	UKP4#	Hadfield *et al*[1]	PRJNA253843	9,001,535	0.107	Human	201
C. parvum parvum	IIaA15G2R1	UKP5#	Hadfield *et al*[1]	PRJNA253845	9,283,240	0.236	Human	201
C. parvum parvum	IIaA15G2R1	UKP6‡	Hadfield. *et al*[1]	PRJNA253846	9,112,937	0.023	Human	201
C. parvum parvum	IIaA17G1R1	UKP7*	Hadfield *et al*[1]	PRJNA253847	9,221,025	0.246	Human	201
C. parvum parvum	IIdA22G1	UKP8*	Hadfield *et al*[1]	PRJNA253848	9,203,336	0.145	Human	201
C. parvum anthroponosum	IIcA5G3p	UKP12*	This study	PRJNA315504	9,325,214	1.686	Human	201
C. parvum anthroponosum	IIcA5G3a	UKP13*	This study	PRJNA315505	9,031,205	1.876	Human	201
C. parvum anthroponosum	IIcA5G3a	UKP14*	This study	PRJNA315506	9,432,159	0.944	Human	201
C. parvum anthroponosum	IIcA5G3a	UKP15‡	This study	PRJNA315507	9,408,807	0.307	Human	201
C. parvum parvum	IIcA5G3j	UKP16‡	This study	PRJNA315508	9,308,724	0.240	Human	201
C. ubiquitum	XIIb	UKUB1*	This study	PRJNA315509	9,060,260	1.812	Human	201
C. ubiquitum	XIIb	UKUB2*	This study	PRJNA315510	9,069,162	0.907	Human	201
C. viatorum	XVaA3f	UKVIA1*	This study	PRJNA492837	11,261,626	0.112	Human	201

* Included in whole genome comparative genomics
‡ Included in whole genome comparative genomics and recombination analysis
# Used only for read mapping onto Iowa II in Figure 3b
¥ Tufts University School of Veterinary Medicine, Medford, Massachusetts (Unpublished genome, CryptoDB.org)

Supplementary Table 2

General Linear Model (GLM) of the pairwise genetic distance (Kxy) between *C. p. parvum* and *C. p. anthroponosum* isolates, with geographic distance as covariate crossed with species. Genetic distances of the gp60 gene between isolates were expressed as Kxy, and these were calculated with the software DnaSP 5.10.1[2]. The geographic distance between isolates (expressed in km as the crow flies) were calculated as the distance between the centre of one country or region to the centre of another using Google Maps (2017). A General Linear Model (GLM) was used to assess differences in the population genetic structure of *C. p. parvum* and *C. p. anthroponosum*. In this model, the pairwise genetic distance (Kxy) was used as the response variable, and species as fixed factor. Species was crossed with geographic distance between sampling points, which was included as a covariate in the model. This interaction term (species x distance) interrogates whether the two regression lines for both species have a similar slope.

Analysis of Variance for Kxy, using Adjusted SS for Tests						
Source	**DF**	**Seq SS**	**Adj SS**	**Adj MS**	**F**	**P**
Spp	1	10.073	9.492	9.492	13.14	0.001
Km	1	18.729	25.359	25.359	35.10	0.000
spp*km	1	34.209	34.209	34.209	47.34	0.000
Error	79	57.082	57.082	0.723		
Total	82	120.093				

Supplementary Table 3

Linear Model of the pairwise genetic distance (Kxy) of *C. p. parvum* isolates versus distance.

Linear Model of Kxy versus geographic distance					
Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	1	52.560	52.560	40.63	0.000
Residual Error	26	33.634	1.294		
Total	27	86.193			

R-Sq. = 61.0%

Supplementary Table 4

Description of positively-selected (>1.0 Ka/Ks) protein-coding genes between *C. parvum parvum* UKP6 and *C. parvum anthroponosum* UKP15.

Chromosome	CryptoDB ID[3]	Nucleotide Diversity (π)[4]	Ka/Ks[5]	Length (bp)	Protein localization[6]
1	cgd1_120	0.0212	2.6169	1293	plas
	cgd1_620	0.0053	1.2919	1314	extr
	cgd1_1230	0.0019	1.5228	3195	extr
	cgd1_1400	0.0017	1.0679	2886	plas
	cgd1_1640	0.0006	1.3054	8628	plas
	cgd1_3760	0.0023	2.3799	3477	nucl
2	cgd2_390	0.0286	1.0009	465	extr
	cgd2_430	0.0099	1.1634	612	extr
	cgd2_940	0.0022	1.5374	3189	plas
	cgd2_2900	0.0118	1.5270	513	cyto
	cgd2_4060	0.0023	1.0462	2163	plas
	cgd2_4370	0.0302	3.2666	1122	extr
3	cgd3_1690	0.0061	1.1526	987	extr
	cgd3_1710	0.0112	2.5777	900	extr
	cgd3_1780	0.0224	2.9534	1653	plas
	cgd3_2080	0.0015	1.0061	3930	plas
	cgd3_3650	0.0035	1.1533	1413	cyto
	cgd3_4180	0.0028	1.1498	3603	extr
4	cgd4_3670	0.0078	1.0479	1404	nucl
	cgd4_3750	0.0052	1.8261	1938	plas
5	cgd5_20	0.0040	2.0821	2280	extr
	cgd5_50	0.0067	1.4633	1962	extr
	cgd5_580	0.0026	1.1164	3405	plas
	cgd5_2560	0.0025	1.5475	2451	nucl
6	cgd6_10	0.1302	1.1110	630	extr
	cgd6_40	0.0334	2.6311	555	extr
	cgd6_640	0.0029	1.6584	2727	cyto
	cgd6_1010	0.0028	1.3601	2145	nucl
	cgd6_3600	0.0054	1.0740	747	nucl
	cgd6_3920	0.0042	2.5510	2367	extr
	cgd6_5110	0.0041	1.0056	8109	plas
	cgd6_5270	0.1194	1.2212	480	extr
	cgd6_5410	0.0052	1.0307	2121	extr
	cgd6_5500	0.0091	1.1566	882	extr
7	cgd7_1280	0.0072	1.2224	561	extr
	cgd7_2600	0.0018	1.8389	4416	cyto
8	cgd8_40	0.0452	1.6397	2652	extr
	cgd8_60	0.0186	1.1000	600	extr
	cgd8_290	0.0043	1.3539	1155	mito
	cgd8_380	0.0045	1.2426	1578	cyto
	cgd8_520	0.0092	1.1355	1101	extr
	cgd8_1570	0.0043	1.3895	1410	mito
	cgd8_2450	0.0019	1.5568	3162	nucl
	cgd8_2550	0.0015	1.5955	3951	plas

Supplementary Table 5

Description of hypervariable (<90.0% amino acid identities) protein-coding genes between *C. parvum parvum* UKP6 (IIaA15G2R1) and *C. hominis* UKH4 (IaA14R3).

Chromosome	CryptoDB ID[3]	% AA IDs[4]	InDel Frameshift[4]	Putative Protein Function[7]	Putative Localization[6]	KaKs[5]
1	cgd1_110	83.4	FS	Secreted protein	extr	0.6075
	cgd1_120	77.2		Uncharacterized	plas	1.0254
	cgd1_130	80.8		IWS1-like protein	plas	0.6841
	cgd1_140	60.7	FS	Predicted secreted protein	extr	0.6060
	cgd1_430	56.0	FS	Uncharacterized	extr	0.4024
	cgd1_470	74.7		Mucin	nucl	0.6481
	cgd1_590	84.7	FS	Proteoglycan/mucin	extr	0.5808
	cgd1_620	88.6		Viral A-type inclusion protein	extr	0.7577
	cgd1_680	73.3		Uncharacterized	nucl	0.3329
	cgd1_900	33.3	FS	Uncharacterized	extr	0.2405
	cgd1_1030	54.9	FS	Uncharacterized	cyto	0.2691
	cgd1_1190	79.4	FS	Uncharacterized	extr	99.000
	cgd1_1320	52.0	FS	Developmental protein	extr	0.0010
	cgd1_1440	88.0	FS	Uncharacterized	cyto	0.2990
	cgd1_1510	89.5	FS	Uncharacterized	mito	0.1463
	cgd1_1650	86.8	FS	Uncharacterized	extr	0.0629
	cgd1_1710	89.2	FS	Phosphoglycerate mutase	mito	0.1439
	cgd1_3290	89.4		Carboxylesterase	plas	0.3342
	cgd1_3430	82.5	FS	Uncharacterized	extr	1.3880
	cgd1_3450	38.5	FS	Uncharacterized	nucl	0.2532
	cgd1_3590	86.2		Membrane associated protein	plas	0.3157
	cgd1_3680	36.1	FS	EGF-like domain protein	plas	0.3002
	cgd1_3850	78.5		Uncharacterized	plas	0.6967
	cgd1_3860	22.1	FS	Deoxyuridine 5'-triphosphate nucleotidohydrolase	mito	0.4853
2	cgd2_390	81.3		Mucin	extr	1.4084
	cgd2_400	82.5		Mucin	extr	0.5439
	cgd2_410	82.9		Mucin	extr	1.0176
	cgd2_420	59.5		Mucin	extr	1.4730
	cgd2_430	71.8		Mucin	extr	0.9262
	cgd2_440	78.5		Mucin	extr	4.1629
	cgd2_450	74.8		Mucin	extr	0.5573
	cgd2_840	84.0	FS	Phosphatidylinositol N-acetylglucosaminyltransferase subunit P	cyto	0.1336
	cgd2_1170	68.0	FS	Zinc finger protein ZPR1	cyto	0.0010
	cgd2_1550	83.8	FS	Origin of replication complex subunit 4	cyto	0.0496
	cgd2_1970	86.3	FS	SAM dependent methyltransferase	nucl	0.2756
	cgd2_2110	73.4	FS	Uncharacterized	plas	0.1043
	cgd2_2180	72.4	FS	Uncharacterized	extr	0.2647
	cgd2_2460	55.2	FS	Insulin growth factor-binding protein	cyto	0.1568
	cgd2_2550	87.9		Lipoprotein	plas	0.4952
	cgd2_2560	66.1	FS	Uncharacterized	extr_plas	0.8179

	cgd2_2570	88.3		Uncharacterized	extr	0.4058
	cgd2_2600	89.6		Uncharacterized	extr	0.4483
	cgd2_2650	84.9	FS	Uncharacterized	plas	0.7622
	cgd2_2900	87.7		Uncharacterized	cyto	0.9958
	cgd2_3140	62.4		Mucin	plas	0.1668
	cgd2_3270	83.1	FS	Phosphoglucomutase/phosphomannomutase family protein	E.R.	0.0661
	cgd2_3280	37.9	FS	Aminopeptidase	plas	0.3054
	cgd2_3370	64.2	FS	Proteasome regulatory subunit Rpn12 family	extr	0.2861
	cgd2_3520	83.2		IWS1 like protein	extr	0.4715
	cgd2_3530	85.7		Eukaryotic translation initiation factor	nucl	0.5232
	cgd2_3610	67.8	FS	WD domain containing protein	extr	0.0719
	cgd2_3780	81.5	FS	Mucin	cyto_nucl	0.1400
	cgd2_3820	52.8	FS	Uncharacterized	extr	0.0010
	cgd2_3970	85.8	FS	RNA recognition family protein	extr	0.1737
	cgd2_4020	89.8		Uncharacterized	extr	0.6982
	cgd2_4310	88.8	FS	Uncharacterized	nucl	0.9573
	cgd2_4370	78.0		Early endosome antigen 1	extr	1.1392
	cgd2_4380	69.9	FS	Mucin	extr	0.7805
3	cgd3_10	83.4		Anchor protein	plas	0.4945
	cgd3_170	72.9	FS	DUF947-domain-containing protein	nucl	0.4624
	cgd3_190	73.6		Mucin	plas	0.1696
	cgd3_370	39.8	FS	Uncharacterized	extr	99.000
	cgd3_630	84.6		Integral membrane protein	plas	0.3972
	Chro.30091	88.6		Proteoglycan	E.R.	0.3263
	cgd3_820	88.7		Uncharacterized	plas	0.8233
	cgd3_1073	52.6	FS	Synaptobrevin family protein	cyto	0.1464
	cgd3_1100	82.4		Nipped-B-like protein	cyto	1.0843
	cgd3_1150	70.2		Uncharacterized	extr	0.7599
	cgd3_1160	85.1		RNA polymerase-associated protein	plas	0.6439
	cgd3_1170	35.7	FS	Uncharacterized	extr	0.3388
	cgd3_1680	58.3	FS	Uncharacterized	plas	0.3542
	cgd3_1690	86.7		Uncharacterized	extr	0.5088
	cgd3_1710	85.7		Uncharacterized	extr	0.9842
	cgd3_1730	87.6		Uncharacterized	extr	1.2875
	cgd3_1740	85.0		Ubiquitin-like protein	mito	0.9198
	cgd3_1750	88.3		Inositol-phosphate phosphatase	extr	0.6631
	cgd3_1760	81.3		Uncharacterized	cyto	0.7370
	cgd3_1770	75.5		Uncharacterized	extr	0.8301
	cgd3_1780	82.2		Antigen	plas	1.1087
	Chro.30271	79.8	FS	Gaa1-like GPI transamidase component	plas	2.3317
	cgd3_2700	88.6	FS	Trafficking protein particle	extr	0.0586
	cgd3_2830	88.4	FS	Uncharacterized	mito	0.1788
	cgd3_4260	87.5		Insulinase like peptidase	plas	0.3161
	cgd3_4270	89.5		Insulinase like peptidase	plas	0.2621
	cgd3_4360	89.0		Uncharacterized	plas	0.4409

4	cgd4_10	86.3		Glutamate receptor	extr	0.6381
	cgd4_32	89.9	FS	Glycoprotein	cyto	0.3236
	cgd4_210	58.3	FS	Ubiquitin-conjugating enzyme 27	cysk	0.0010
	cgd4_770	88.1		Trichohyalin	cyto	0.1082
	cgd4_920	75.5	FS	Histidine phosphatase superfamily	plas	0.1992
	cgd4_1000	11.2		Cell wall anchor protein	nucl	99.000
	cgd4_1280	74.3	FS	Rtf2 RING-finger family protein	mito	0.5351
	cgd4_1300	58.0		Mucin	nucl	0.4326
	cgd4_2160	42.9	FS	Ribonuclease	extr_plas	0.3239
	cgd4_2450	86.6	FS	Tubulin-specific chaperone C	cyto	0.4707
	cgd4_2500	87.4	FS	Uncharacterized	extr	0.2272
	cgd4_2510	81.6	FS	Uncharacterized	extr	0.7576
	cgd4_2760	56.8	FS	Mitotic-spindle organizing protein	mito	0.3642
	cgd4_2830	87.6	FS	Mra1/NEP1 like protein	extr	0.2191
	cgd4_3060	60.8	FS	Uncharacterized	cyto	0.9623
	cgd4_3350	63.8	FS	Mob1/phocein family protein	extr	0.5179
	cgd4_3520	88.1		Proteophosphoglycan	nucl	0.2599
	cgd4_3550	85.4		Kazal-type serine protease inhibitor domain-containing protein	extr	0.2612
	cgd4_3630	65.7		Cross-beta structure silk protein 1	nucl	0.6389
	cgd4_3640	77.0	FS	Uncharacterized	cyto	0.3190
	cgd4_3650	55.9	FS	Uncharacterized	extr	1.3451
	cgd4_3660	37.8	FS	Uncharacterized	cyto	1.1696
	cgd4_3670	75.1		Collagen-like protein	nucl	0.4561
	cgd4_3680	87.8		Uncharacterized	cyto	0.5627
	cgd4_3690	70.2		Glycine-rich cell wall structural protein	plas	0.4939
	cgd4_3930	74.4	FS	Exosome complex component	mito	0.0233
	cgd4_3970	82.8		GPI-anchored protein	plas	0.2715
	cgd4_4070	30.8	FS	Uncharacterized	extr	0.6830
	cgd4_4210	56.5	FS	Antigen	plas	0.1748
	cgd4_4253	80.1	FS	Uncharacterized	cyto	0.3880
	cgd4_4390	70.6	FS	Uncharacterized	mito	0.0556
	cgd4_4470	88.4		Dentin sialophosphoprotein	plas	0.3753
	cgd4_4480	89.0		Uncharacterized	plas	0.4933
	cgd4_4500	73.4	FS	Proteophosphoglycan	nucl	0.7551
5	cgd6_5500	64.3	FS	Uncharacterized	cyto	0.2680
	cgd5_10	87.5		S-antigen protein	extr	0.6797
	cgd5_20	89.0		GPI-anchored adhesin-like	extr	0.5210
	Cgd5_40	71.8		Erythrocyte membrane protein	extr_plas	0.9587
	cgd5_50	82.1		Uncharacterized	extr_plas	1.0310
	cgd5_130	89.4		Ferlin like type II membrane associated protein	plas	0.0691
	cgd5_450	89.0		Putative RING zinc finger	nucl	0.1065
	cgd5_1090	87.9	FS	Uncharacterized	extr	0.6167
	cgd5_1580	84.8	FS	Uncharacterized	cyto	0.6614
	cgd5_1940	89.6		Viral A-type inclusion protein	nucl	0.3923
	cgd5_2180	81.8		Mucin 17-like protein	nucl	0.1433

	cgd5_2960	85.1	FS	Putative U5 small nuclear ribonucleoprotein 200 kDa helicase	plas	0.2925
	cgd5_3190	86.2	FS	Protein kinase domain protein	cyto	0.1631
	cgd5_3440	56.1	FS	Uncharacterized	extr	0.5517
	cgd5_3490	89.9		Biotin-protein ligase	extr	0.4294
	Chro.50010	44.5	FS	Proteophosphoglycan	plas	0.3678
6	cgd6_10	46.6		Proteophosphoglycan	extr	0.2680
	cgd6_40	72.5		Antigen	extr	0.6395
	cgd6_50	36.6	FS	Uncharacterized	extr	0.9261
	cgd6_60	88.1		Protease	nucl	0.3409
	cgd6_170	82.6	FS	Synaptobrevin-like protein	cyto	0.3476
	cgd6_260	57.7	FS	Diacylglycerol acyltransferase	plas	0.0922
	cgd6_340	63.1	FS	Uncharacterized	extr	0.6149
	cgd6_780	86.9	FS	Sporozoite cysteine-rich protein	plas	0.2063
	cgd6_920	48.8	FS	26S proteasome regulatory subunit 8	cyto	99.000
	cgd6_960	74.8	FS	Cysteinyl-tRNA synthetase	cyto_nucl	0.0802
	cgd6_1080	69.2		Glycoprotein	extr	0.5341
	cgd6_1170	89.7		Uncharacterized	cyto	0.5669
	cgd6_1620	89.2	FS	Uncharacterized	cyto	0.4667
	cgd6_2130	80.8	FS	RNA methyltransferase	plas	0.4113
	cgd6_2140	48.9	FS	Ion channel protein	cyto	0.2336
	cgd6_2270	47.6	FS	Membrane-associated protein	plas	0.1257
	cgd6_2500	77.8	FS	Rhoptry protein	plas	0.1178
	cgd6_2660	75.2	FS	DNA repair helicase	nucl	0.1060
	cgd6_2800	86.4	FS	Ras-related GTP-binding protein	cysk	0.1877
	cgd6_3050	81.8		Mucin	extr	0.7440
	cgd6_3360	71.8	FS	FYVE and coiled-coil domain-containing protein	extr	0.1301
	cgd6_3770	88.6	FS	Insulin-degrading enzyme	cyto	0.0661
	cgd6_3930	81.5		Glycoprotein	nucl	0.5344
	cgd6_3940	71.2		Glycoprotein	mito	1.8627
	cgd6_4100	45.5	FS	Uncharacterized	extr	0.2559
	cgd6_4230	89.0		Cement protein 3B	extr	0.5390
	cgd6_4670	56.2	FS	Splicing factor 3A subunit 3	cyto	0.1523
	cgd6_4740	84.1		Transmembrane protein 64	plas	0.4542
	cgd6_4980	46.5	FS	Uncharacterized	plas	99.000
	cgd6_5110	86.4	FS	Reticulocyte binding protein	plas	0.3134
	cgd6_5270	88.8		Uncharacterized	extr	0.3736
	cgd6_5400	70.5		Mucin	extr	0.1940
	cgd6_5410	85.7		Mucin	extr	0.3327
	cgd6_5430	86.6		GPI-anchored adhesin-like protein	plas	0.6845
7	cgd5_4530	23.2	FS	Uncharacterized	E.R._mito	0.1754
	cgd7_10	81.1	FS	Binding protein	plas	0.5083
	cgd7_1210	88.8		Integral membrane protein	extr	1.3153
	cgd7_1280	76.7		Glycoprotein	extr	0.6014
	cgd7_1370	89.1		Uncharacterized	extr	0.6406
	cgd7_1870	87.6	FS	Uncharacterized	extr	1.2958

	cgd7_2120	67.9	FS	Uncharacterized	extr	0.3770
	cgd7_2350	48.6	FS	Uncharacterized	plas	1.2958
	cgd7_2870	83.0	FS	Titin	nucl	0.5227
	cgd7_3420	30.7	FS	Uncharacterized	mito	0.1356
	cgd7_3440	68.2	FS	Uncharacterized	cyto	99.000
	cgd7_3800	82.1	FS	Uncharacterized	extr	0.1719
	cgd7_4020	88.8		Mucin	plas	0.0632
	cgd7_4260	89.4	FS	Uncharacterized	nucl	0.1824
	cgd7_4300	51.7	FS	Zinc finger, C2H2 type domain	cyto	0.1198
	cgd7_4310	82.7	FS	Cysteine-rich secretory protein	extr	0.1025
	cgd7_4430	83.3		Glycosyl transferase family	extr	0.6875
	cgd7_4500	81.9		Proteoglycan/glycoprotein	extr	0.6241
	cgd7_5400	85.1	FS	Uncharacterized	extr	0.0897
	cgd7_5510	89.1		Chromosome partition protein Smc	extr	0.9346
	cgd7_5520	82.5		Glycoprotein	mito	0.2623
8	cgd8_10	86.8		Uncharacterized	cyto	0.4493
	cgd8_20	86.8		Uncharacterized	plas	0.4105
	cgd8_30	87.3	FS	Uncharacterized	nucl	0.4544
	cgd8_40	79.3		Uncharacterized	plas	0.9231
	cgd8_50	89.6		Uncharacterized	plas	0.3737
	cgd8_60	71.3	FS	Uncharacterized	extr	0.7822
	cgd8_520	83.0		Histone H5	extr	0.6743
	cgd8_660	72.4	FS	Mucin	E.R.	0.6837
	cgd8_700	87.0		Mucin	plas	0.3744
	cgd8_1020	74.3	FS	N terminus of Rad21/Rec8 like protein	cyto	0.1515
	cgd8_1160	89.8		Mucin	plas	0.1597
	cgd8_1220	80.9	FS	Mucin	cyto	0.2073
	cgd8_1410	75.7	FS	DNA primase large subunit	cyto	0.0386
	cgd8_1570	71.6	FS	CCCH like finger domain nucleoporin	mito	0.2653
	cgd8_1750	89.7		Uncharacterized	extr	0.2194
	cgd8_1770	89.7		Proteophosphoglycan	plas	0.2210
	cgd8_1820	57.1	FS	Uncharacterized	extr	1.8643
	cgd8_2140	52.9	FS	Uncharacterized	plas	0.2520
	cgd8_2160	84.6		Poly(ADP-ribose) glycohydrolase	plas	0.6824
	cgd8_2220	85.8	FS	Male gamete fusion factor family	nucl	0.1903
	cgd8_2240	84.1	FS	Histidine phosphatase superfamily	cyto	4.2459
	cgd8_2590	58.3	FS	Uncharacterized	plas	0.0862
	cgd8_2800	63.5	FS	Mucin	plas	0.2877
	cgd8_3120	86.6	FS	Uncharacterized	extr	0.6230
	cgd8_3200	86.7	FS	Ubiquitin carboxyl-terminal hydrolase	cyto	0.2506
	cgd8_3540	89.9	FS	Uncharacterized	plas	0.5655
	cgd8_3550	38.5	FS	Uncharacterized	cyto	0.1517
	cgd8_3550	40.5	FS	Uncharacterized	mito	1.8383
	cgd8_3650	72.2	FS	Trafficking protein particle complex	cysk	0.3312
	cgd8_3670	85.3	FS	Uncharacterized	mito	0.1203

	cgd8_4190	73.7		Mucin	cyto	0.3961
	cgd8_4480	88.8		Type VI secretion system Vgr family	nucl	0.1217
	cgd8_4550	76.6	FS	Uncharacterized	cyto	0.3963
	cgd8_4740	66.2	FS	Phosphopantetheinyl transferase	cyto	0.1954
	cgd8_4820	23.6	FS	Transcription initiation factor IID	cyto	0.4455
	cgd8_4860	89.8	FS	Antigen	extr	0.3749
	cgd8_5050	70.7	FS	Palmitoyltransferase	plas	0.4310
	cgd8_5290	89.1		Glycoprotein	plas	0.3766
	cgd8_5360	26.9	FS	Glycoprotein	extr	0.7168
	cgd8_5370	64.6		Uncharacterized	extr	1.9001
	cgd8_5380	75.0		Rap guanine nucleotide exchange factor	extr	0.9647
	cgd8_5390	88.4		Uncharacterized	extr	1.2072
	cgd8_5420	24.7	FS	Uncharacterized	extr	0.5848

Supplementary Table 6

Description of hypervariable (<90.0% amino acid identities) protein-coding genes between *C. parvum parvum* UKP6 (IIaA15G2R1) and *C. parvum anthroponosum* UKP15 (IIcA5G3a).

Chromosome	CryptoDB ID[3]	% AA IDs[4]	InDel Frameshift[4]	Putative Protein Function[7]	Putative Localization[6]	KaKs[5]
1	cgd1_150	25.8	FS	Autophagy-related protein 11	plas	0.3287
	cgd1_470	80.3		Mucin	nucl	0.6767
2	cgd2_3140	85.4		Mucin	plas	0.1458
	cgd2_3530	87.8		Eukaryotic translation initiation factor	nucl	2.2698
3	cgd3_370	26.0	FS	Uncharacterized	extr	0.0010
	cgd3_1150	89.8		Uncharacterized	extr	0.7918
	cgd3_1160	38.2	FS	RNA polymerase-associated protein	plas	1.1871
	cgd3_1170	82.1	FS	Uncharacterized	extr	0.6709
	cgd3_1680	65.3	FS	Uncharacterized	plas	0.0010
4	cgd4_1280	74.3	FS	Rtf2 RING-finger family protein	mito	1.3363
	cgd4_1300	79.8		Mucin	nucl	0.5568
	cgd4_3690	44.2	FS	Glycine-rich cell wall structural protein	plas	1.2300
	cgd4_3660	40.1	FS	Uncharacterized	cyto	0.7244
	cgd4_3060	36.9	FS	Uncharacterized	cyto	0.0010
	cgd4_2830	89.7	FS	Mra1/NEP1 like protein	extr	0.0010
	cgd4_4070	31.2	FS	Uncharacterized	extr	1.6923
	cgd4_4390	71.3	FS	Uncharacterized	mito	0.2986
	cgd4_4470	29.4	FS	Dentin sialophosphoprotein	plas	0.3275
	cgd4_4500	67.8	FS	Proteophosphoglycan	nucl	0.8056
5	Cgd5_40	81.9		Erythrocyte membrane protein	extr_plas	0.4943
	cgd5_1670	84.4	FS	Lysine-rich arabinogalactan protein	mito	0.0010
	cgd5_2180	86.8		Mucin 17-like protein	nucl	0.2466
	Chro.50010	77.3		Proteophosphoglycan	plas	0.3196
6	cgd6_10	66.8		Proteophosphoglycan	extr	1.1110
	cgd6_40	89.1		Antigen	extr	0.7261
	cgd6_50	44.5	FS	Uncharacterized	extr	99.0000
	cgd6_170	89.9	FS	Synaptobrevin-like protein	cyto	99.0000
	cgd6_250	83.4	FS	TatD-like deoxyribonuclease	cyto	0.4839
	cgd6_340	60.9	FS	Uncharacterized	extr	0.7206
	cgd6_520	89.4		Ser/Thr protein kinase	cyto	0.1471
	cgd6_780	86.6	FS	Sporozoite cysteine-rich protein	plas	0.2870
	cgd6_1080	70.4		Glycoprotein	extr	0.6763
	cgd6_5270	79.4		Uncharacterized	extr	1.2639
7	cgd7_2120	63.6	FS	Uncharacterized	extr	1.1248
	cgd7_4310	83.4	FS	Cysteine-rich secretory protein	extr	0.4982
8	cgd8_10	75.2		Uncharacterized	cyto	0.5436
	cgd8_20	81.0		Uncharacterized	plas	0.9078
	cgd8_30	85.8		Uncharacterized	nucl	0.8820
	cgd8_40	89.4		Uncharacterized	plas	1.4514
	cgd8_1570	71.6		CCCH like finger domain nucleoporin	mito	1.3897

	cgd8_4190	87.0		Mucin	cyto	0.6159
	cgd8_4550	78.4	FS	Uncharacterized	cyto	0.5160
	cgd8_5190	85.9		BRCA2 family protein	plas	0.5171
	cgd8_5420	78.8	FS	Uncharacterized	extr	0.5314

Supplementary Table 7

Summary of RDP4[8] recombination results with position of breakpoints, and estimated dates of divergence (thousands of generations ago) between the sequences that are related to the sequences involved in the genetic exchange. The HybridCheck[9] algorithm was used to estimate the divergence time of the recombinant blocks identified by RDP4. The "major parent" is related to the greater part of the recombinant's sequence (i.e. it is generally the recipient). The "minor parent" is related to the sequences in the proposed recombinant region (i.e. the donor). For the analysis n=4: *C. p. parvum* subtypes IIaA15G2R1 (UKP6; IIa) and IIcA5G3j (UKP16; IIc-j), *C. p. anthroponosum* subtype IIcA5G3a (UKP15; IIc-a), and *C. hominis* subtype IbA10G2 (UKH1; Ib). Subtyping was based on gp60 genotyping. The p-value represents the probability that the identified recombination block is the result of the accumulation of mutations rather than by recombination. The critical value is Bonferroni corrected, α'=0.05/n, with n equal to the number of recombination events detected.

Breakpoints (bp)		Recombinant	Major parent	Minor Parent	RDP *p*-value	CDSs encoded within	Divergence Dating (TGA)
CHROMOSOME 1							
82251	**104422**	IIa	IIc-j	Unknown	8.28E-240	cgd1_370 - cgd1_490	NA
82251	**93181**	Ib	IIc-a/IIc-j	Unknown	4.46E-08	cgd1_370 - cgd1_430	NA
100170	**100278**	IIa	IIc-a/IIc-j	Unknown	3.07E-03	cgd1_470	NA
100631	**100831**	Ib	Unknown	IIa	5.25E-04		NA
109846	**110180**	Ib	Unknown	IIc-a	8.34E-13	Intergenic cgd1_510 - cgd1_520	NA
111232	**111726**	IIc-a	IIa/IIc-j	Ib	1.26E-13	cgd1_530	32358 (95% CI: 24014-42302)
115061	**116161**	IIc-j/IIa	IIc-a	Ib	2.78E-07	cgd1_550	8476 (95% CI: 5665-12074)
127173	**136648**	IIa	IIc-j	IIc-a	2.93E-72	cgd1_580 - cgd1_590	8234 (95% CI: 7177-9388)
136649	**140781**	IIc-j	IIa	IIc-a	1.18E-15	cgd1_590 - cgd1_600	6513 (95% CI: 5166-8073)
142478	**150610**	IIa	IIc-j	IIc-a	8.96E-16	cgd1_610 - cgd1_640	3738 (95% CI: 3006-4580)
376602	**386949**	IIc-a	Unknown	IIc-j/IIa	2.92E-02	cgd1_1580 - cgd1_1640	NA
734690	**744935**	IIa	IIc-j	IIc-a	4.18E-04	cgd1_3290 - cgd1_3340	1403 (95% CI: 1016-1878)
CHROMOSOME 2							
53785	**55454**	IIc-a	IIa/IIc-j	Ib	1.38E-14	cgd2_160	13159 (95%CI: 10150-16693)
57056	**57358**	IIc-a	IIa/IIc-j	Ib	1.54E-06	cgd2_140	28412 (95% CI: 18652-40881)
58483	**58812**	IIc-a	IIa/IIc-j	Unknown	3.16E-08	cgd2_120	NA
61997	**62206**	IIc-a	IIa/IIc-j	Ib	1.78E-09	Intergenic cgd2_110 - cgd2_100	39623 (95% CI: 26118-56711)
64582	**65341**	IIc-a	IIa/IIc-j	Ib	8.90E-11	cgd2_90	13627 (95% CI: 9320-19056)
67242	**67933**	IIc-a	IIa/IIc-j	Ib	2.23E-10		19260 (95% CI: 13841-25877)
71503	**72990**	IIc-a/IIa	IIc-j	Ib	2.82E-19	cgd2_80	13490 (95% CI: 10302-17256)
75512	**76343**	IIc-a	IIa/IIc-j	Ib	5.18E-08	cgd2_70	10614 (95% CI: 7015-15246)
79931	**80238**	IIc-a	IIa/IIc-j	Ib	4.95E-10	Intergenic cgd2_70 - cgd2_60	27117 (95% CI: 17794-39042)
294024	**294928**	IIa	IIc-j	Unknown	4.32E-10	cgd2_1370	NA
341750	**405045**	IIc-a	Unknown	IIc-j	4.28E-06	cgd2_1690 - cgd2_2040	NA
432700	**506795**	IIc-j	IIa	Unknown	6.55E-04	cgd2_2170 - cgd2_2560	NA
625528	**632428**	IIa	IIc-j	Unknown	1.93E-06	cgd2_3080 - cgd2_3110	NA
CHROMOSOME 3							
220866	**220932**	IIc-a	IIc-j/IIa	Unknown	4.15E-08	cgd3_720	NA
272798	**279815**	IIc-j	IIa	IIc-a	5.11E-04	cgd3_920 - cgd3_960	1335 (95% CI: 890-1907)
319189	**319570**	IIc-a	IIc-j/IIa	Unknown	9.41E-05	cgd3_1150	NA
321883	**322660**	IIc-j	IIa	IIc-a	1.75E-15	cgd3_1160	23555 (95% CI: 17832-30326)
797968	**799943**	IIc-a	IIc-j/IIa	Ib	1.43E-77	cgd3_3370	23504 (95% CI: 19790-27632)
995078	**1030425**	IIc-j	IIa	IIc-a	2.18E-17	cgd3_4190 - cgd3_4280	776 (95% CI: 616-961)

CHROMOSOME 4

3370	**5132**	IIa/IIc-j	IIc-a	Ib	1.70E-24	cgd4_20	21044 (95% CI: 17332-25227)
5137	**5788**	IIc-a	IIa/IIc-j	Ib	5.46E-41		106298 (95% CI: 93209-120222)
848234	**849840**	IIc-j	IIa	IIc-a	2.06E-13	cgd4_3630	10881 (95% CI: 7944-14449)
865724	**865737**	IIc-a	IIa/IIc-j	Ib	1.89E-06	cgd4_3690	34157 (95% CI: 27128-42213)
1054213	**1054636**	IIc-a	IIc-j/IIa	Ib	4.80E-14	cgd4_4480	40826 (95% CI: 30656-52820)
1057053	**1058582**	IIc-j/IIa	IIc-a	Ib	1.27E-54	cgd4_4490	35621 (95% CI: 30459-41290)
1058583	**1058932**	IIc-a	IIa/IIc-j	Ib	2.10E-13	Intergenic cgd4_4490 - cgd4_4500	37164 (95% CI: 26671-49871)
1059044	**1059293**	IIc-j/IIa	IIc-a	Ib	5.69E-12	cgd4_4500	55261 (95% CI: 40390-72926)
1059418	**1060146**	IIc-a	IIa/IIc-j	Unknown	3.67E-47		NA
1060336	**1060469**	IIa/IIc-j	IIc-a	Ib	3.22E-07		62039 (95% CI: 41237-87846)
1060678	**1060737**	IIc-a	IIa/IIc-j	Ib	1.45E-06		146415 (95% CI: 101853-197080)
1061059	**1061153**	IIc-a	IIa/IIc-j	Ib	2.89E-04		79429 (95% CI: 51637-113811)
1061156	**1061888**	IIc-j/IIa	IIc-a	Ib	1.65E-77	Intergenic cgd4_4500 - 3' telomere	43165 (95% CI: 35048-52324)
1061941	**1062512**	IIc-a	IIc-j/IIa	Unknown	5.83E-27		NA
1062847	**1063606**	IIc-j/IIa	IIc-a	Ib	1.97E-61		75905 (95% CI: 65591-87050)

CHROMOSOME 5

3694	**6176**	IIa	IIc-j	IIc-a	1.03E-23	Chro.50010	9774 (95% CI: 7624-12287)
585260	**586337**	IIc-j	IIc-a/IIa	Ib	3.76E-28	cgd5_2180	81221 (95% CI: 71131-92033)
649071	**649362**	Ib	Unknown	IIa/IIc-j	8.09E-51	cgd5_1940	NA
1031972	**1033136**	IIc-j/IIa	IIc-a	Ib	2.15E-45	cgd5_40	62872 (95% CI: 52872-73873)

CHROMOSOME 6

49	**140**	IIc-j	IIa	Unknown	3.56E-05	Chro.60010	NA
146	**1792**	IIc-j	IIa	Unknown	2.59E-164		NA
1793	**1905**	IIc-j	IIa	Ib	1.43E-12	Intergenic Chro.60010 - cgd6_10	100894 (95% CI: 72306-134229)
1986	**2351**	IIc-j	IIa	Unknown	7.34E-43		NA
2352	**2537**	IIc-j	IIa	Ib	8.61E-23		108012 (95% CI: 81123-138605)
2538	**2963**	IIc-j	IIa	Unknown	3.91E-09		NA
3510	**3670**	IIc-a	IIa	Ib	3.10E-02		33270 (95% CI: 19516-51854)
4026	**6334**	IIa	IIc-j	IIc-a	6.78E-144	cgd6_10	6820 (95% CI: 4794-9277)
7166	**7713**	IIc-j/IIa	IIc-a	Ib	5.31E-18	Intergenic cgd6_10 - cgd6_20	51444 (95% CI: 41285-62947)
7784	**7896**	IIa/IIc-j	IIc-a	Ib	2.21E-04		73535 (95% CI: 49092-103541)
8033	**8972**	Ib	Unknown	IIc-a	2.03E-14	cgd6_20	NA
9758	**9992**	IIc-a	IIa	Ib	1.12E-03		25112 (95% CI: 15057-38631)
10386	**12685**	IIc-j	IIa	IIc-a	3.16E-32	cgd6_30 - cgd6_40	8573 (95% CI: 6516-11016)
13148	**13482**	IIc-a	IIc-j/IIa	Unknown	7.84E-06	Intergenic cgd6_40 - cgd6_50	NA
14883	**18178**	IIa	IIc-j	IIc-a	2.09E-24	cgd6_50	5770 (95% CI: 4366-7444)
20061	**20401**	IIc-a	IIa/IIc-j	Unknown	3.04E-19	cgd6_60	NA
20936	**21391**	IIc-j/IIa	IIc-a	Ib	7.46E-14		54420 (95% CI: 43213-67169)
186255	**187077**	IIa/IIc-j	IIc-a	Ib	1.44E-06	cgd6_800	10123 (95% CI: 6602-14690)
240190	**240717**	IIa/IIc-j	IIc-a	Ib	3.02E-06	cgd6_1020	16881 (95% CI: 11181-24180)
245902	**247871**	IIc-a	Unknown	IIa	2.52E-31	cgd6_1060	NA
247872	**256568**	IIa	IIc-j	Unknown	2.04E-228	cgd6_1060 - cgd6_1100	NA
1225101	**1225478**	IIc-a	IIa/IIc-j	Ib	1.49E-04	cgd6_5260	19386 (95% CI: 12309-28648)
1226191	**1226342**	IIc-a	IIc-j/IIa	Unknown	1.28E-15	cgd6_5260 - cgd6_5270	NA
1226343	**1226614**	IIc-a	IIa/IIc-j	Ib	8.40E-13	cgd6_5270	45290 (95% CI: 32355-60923)

1276817	**1278061**	IIc-a	IIa	Ib	7.88E-28	cgd6_5450	22826 (95% CI: 18282-28028)
1278062	**1278345**	IIc-a	IIa	Unknown	5.11E-07	Intergenic cgd6_5450 - cgd6_5500	NA
1278346	**1280578**	IIc-a	IIa	Ib	4.13E-90	cgd6_5500	109055 (95% CI: 95467-123511)

CHROMOSOME 7

285016	**292270**	IIc-j	IIa	Unknown	3.04E-06	cgd7_1150 - cgd7_1170	NA
317243	**319588**	IIc-j	IIa	IIc-a	3.97E-46	cgd7_1270	15486 (95% CI: 12724-18608)
878265	**897621**	IIc-j	IIa	IIc-a	2.69E-08	cgd7_3910 - cgd7_4020	820 (95% CI: 603-1083)
897622	**898690**	IIc-j	IIa	IIc-a	7.11E-24	cgd7_4020	7422 (95% CI: 3670-13102)
897728	**898242**	IIa	IIc-j	Unknown	1.07E-62		NA
898691	**899005**	Ib	Unknown	IIc-j/IIa	2.16E-14		NA
899011	**935740**	IIc-j	IIa	IIc-a	5.23E-25	cgd7_4020 - cgd7_4220	1241 (95% CI: 1039-1467)
1055570	**1063864**	IIc-j	IIa	IIc-a	1.45E-03	cgd7_4710 - cgd7_4750	887 (95% CI: 560-1323)

CHROMOSOME 8

80	**1150**	IIc-a	IIa/IIc-j	Unknown	9.69E-75	cgd8_10	NA
1334	**1408**	IIc-a	IIa/IIc-j	Unknown	2.72E-08	Intergenic cgd8_10 - cgd8_20	NA
1409	**1526**	IIc-a	IIa/IIc-j	Ib	6.92E-07		79550 (95% CI: 54516-109778)
3201	**3369**	IIc-a	IIa/IIc-j	Unknown	1.18E-08	cgd8_20	NA
3623	**5676**	IIc-j/IIa	IIc-a	Ib	1.29E-110		57629 (95% CI: 51933-63671)
5677	**5972**	IIc-a	IIc-j/IIa	Ib	7.91E-06		33185 (95% CI: 22613-46384)
6026	**7033**	IIc-j/IIa	IIc-a	Ib	6.26E-40	cgd8_30	43724 (95% CI: 36564-51671)
7274	**9938**	IIc-a/IIc-j	IIa	Unknown	2.95E-78		NA
10005	**11970**	IIc-j	IIa	IIc-a	6.41E-07	cgd8_40	7579 (95% CI: 5497-10123)
12805	**14933**	IIc-j	IIa	IIc-a	9.67E-17	cgd8_40 - cgd8_50	20625 (95% CI: 17271-24367)
15040	**26389**	IIa	IIc-j	IIc-a	4.29E-19	cgd8_50 - cgd8_100	3927 (95% CI: 3283-4650)
42714	**48676**	IIa	IIc-j	IIc-a	6.45E-14	cgd8_170 - cgd8_180	2321 (95% CI: 1671-3121)
75004	**84938**	IIc-j	IIa	Unknown	4.04E-06	cgd8_300 - cgd8_350	NA
547697	**563658**	IIa	IIc-j	IIc-a	5.21E-33	cgd8_2090 - cgd8_2150	3327 (95% CI: 2824-3886)
563659	**564762**	IIc-j	IIa	IIc-a	2.16E-26	cgd8_2160	20224 (95% CI: 15722-25475)
564902	**618348**	IIa	IIc-j	IIc-a	1.76E-115	cgd8_2160 - cgd8_2400	3106 (95% CI: 2834-3395)
584382	**584669**	Ib	Unknown	IIc-j	5.59E-04	cgd8_2260	NA
618349	**628553**	IIa	IIc-j	Unknown	2.73E-08	cgd8_2400 - cgd8_2440	NA
1085940	**1086106**	IIc-a	IIa/IIc-j	Unknown	9.40E-32	cgd8_4480	NA

Supplementary Table 8

Whole genome comparison of two outbreak strain WGS reveals estimated mutation accumulation rates per generation for *Cryptosporidium spp*.

UKP4 v UKP6 Whole Genome Comparison	
Sampling separation	7 days
No. of sites in WGA (bp)	9086411
No. of SNPs	10
Nucleotide diversity	0.0000011
No. of indel sites	78
No. of indel events	35
Total no. of polymorphisms (SNPs + indel Events)	45
Per base SNP mutation rate per generation (μ)	9.50E-08
Per base indel rate per generation (μ)	3.32E-07
Combined mutation rate per generation (μ)	4.27E-07

Supplementary Table 9

Oocyst infectivity and intensity rates in human volunteers summarized from peer-reviewed publications.

Reference	Challenge organism	Challenge dose	Onset of Excretion (days)	Duration of Excretion (days)	Total no. of oocysts excreted	Estimated no. of oocyst generations	Estimated no. of days/generation
10	*C. parvum*	100	7.5	3.5	1.8×10^6	4-5	2-4
	C. parvum	300	5	3	3.5×10^6	3-4	2-3
	C. parvum	1,000	4	11	3.1×10^8	4-5	3-4
	C. parvum	3,000	5	6	2.1×10^7	~3	3-4
11	*C. meleagridis*	10,000	8	3	4.5×10^8	~3	3-4

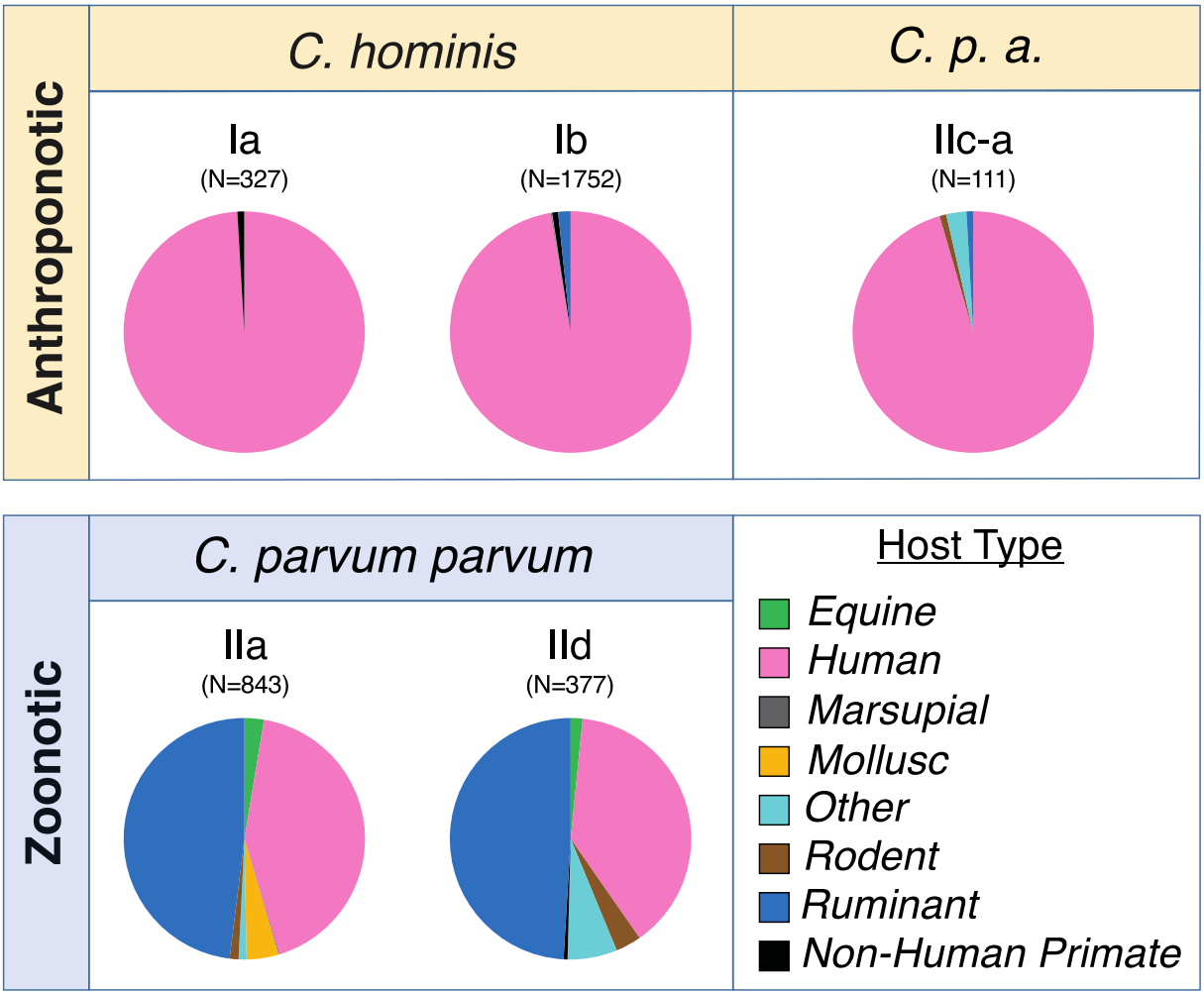
Supplementary Table 10

Description of neutrally-evolving (Ka/Ks = 0.2-0.6; 93.0-98.0% nucleotide IDs) protein-coding genes between *C. parvum parvum* UKP6 and *C. hominis* UKH4 used in the concatenated phylogeny.

Chromosome	CryptoDB ID (*C. hominis*)	CryptoDB ID (*C. parvum*)	Ka/Ks	% Nuc Ids
1	Chro.10076	cgd1_640	0.319577	96.6
	Chro.10167	cgd1_1450	0.438804	96.73
	Chro.10199	cgd1_1730	0.569446	95.58
	Chro.10229	cgd1_2000	0.564612	96.80
	Chro.10411	cgd1_3650	0.497442	95.94
	Chro.10424	cgd1_3780	0.511207	95.87
	Chro.10425	cgd1_3790	0.346492	96.8
2	Chro.20024	cgd2_180	0.4812	96.2
	Chro.20105	cgd2_940	0.382475	96.0
	Chro.20262	cgd2_2470	0.314043	95.7
	Chro.20223	cgd2_2060	0.361484	97.6
	Chro.20326	cgd2_3110	0.31982	95.32
	Chro.20388	cgd2_3630	0.586577	96.30
	Chro.20406	cgd2_3810	0.33444	97.90
3	Chro.30055	cgd3_380	0.386803	96.18
	Chro.30132	cgd3_1010	0.390058	96.03
	Chro.30206	cgd3_1720	0.407783	96.09
	Chro.30299	cgd3_2600	0.366692	97.25
	Chro.30349	cgd3_3070	0.326581	96.76
	Chro.30377	cgd3_3310	0.511435	95.60
	Chro.30413	cgd3_3650	0.262038	97.25
	Chro.30476	cgd3_4230	0.333963	96.12
4	Chro.40051	cgd4_370	0.111926	97.55
	Chro.40248	cgd4_2180	0.387906	97.82
	Chro.40252	cgd4_2210	0.217421	97.63
	Chro.40294	cgd4_2620	0.466828	96.92
	Chro.40317	cgd4_2820	0.504021	96.98
	Chro.40433	cgd4_3800	0.509732	97.39
	Chro.40495	cgd4_4360	0.341557	96.46
	Chro.40503	cgd4_4440	0.350652	97.20
5	Chro.50012	cgd5_3600	0.292362	96.80
	Chro.50084	cgd5_2890	0.425943	96.54
	Chro.50103	cgd5_2730	0.410499	97.23
	Chro.50107	cgd5_2700	0.527435	96.40
	Chro.50155	cgd5_2250	0.249098	96.80
	Chro.50195	cgd5_1860	0.389703	96.68
	Chro.50250	cgd5_1340	0.416003	97.1
	Chro.50420	cgd5_4240	0.322667	96.63
6	Chro.60245	cgd6_2100	0.313076	97.4
	Chro.60295	cgd6_2560	0.382122	96.83
	Chro.60314	cgd6_2720	0.462682	96.05
	Chro.60470	cgd6_4090	0.36524	96.51
	Chro.60490	cgd6_4280	0.366079	96.13
	Chro.60610	cgd6_5300	0.4904	97.43
	Chro.60619	cgd6_5370	0.441644	96.72
7	Chro.70047	cgd7_340	0.333681	96.19
	Chro.70111	cgd7_890	0.318737	96.0
	Chro.70152	cgd7_1270	0.484978	95.8
	Chro.70160	cgd7_1330	0.419706	96.1
	Chro.70211	cgd7_1810	0.292609	96.72
	Chro.70267	cgd7_2340	0.297261	96.8
	Chro.70296	cgd7_2600	0.318605	96.4
	Chro.70395	cgd7_3550	0.500737	96.76
8	Chro.80024	cgd8_140	0.366147	97.36
	Chro.80102	cgd8_830	0.505411	96.50
	Chro.80229	cgd8_1960	0.378299	96.38
	Chro.80245	cgd8_2080	0.435382	96.39
	Chro.80332	cgd8_2850	0.437901	96.7
	Chro.80353	cgd8_3030	0.287705	96.45
	Chro.80409	cgd8_3560	0.438279	96.96
	Chro.80605	cgd8_5310	0.470142	96.32

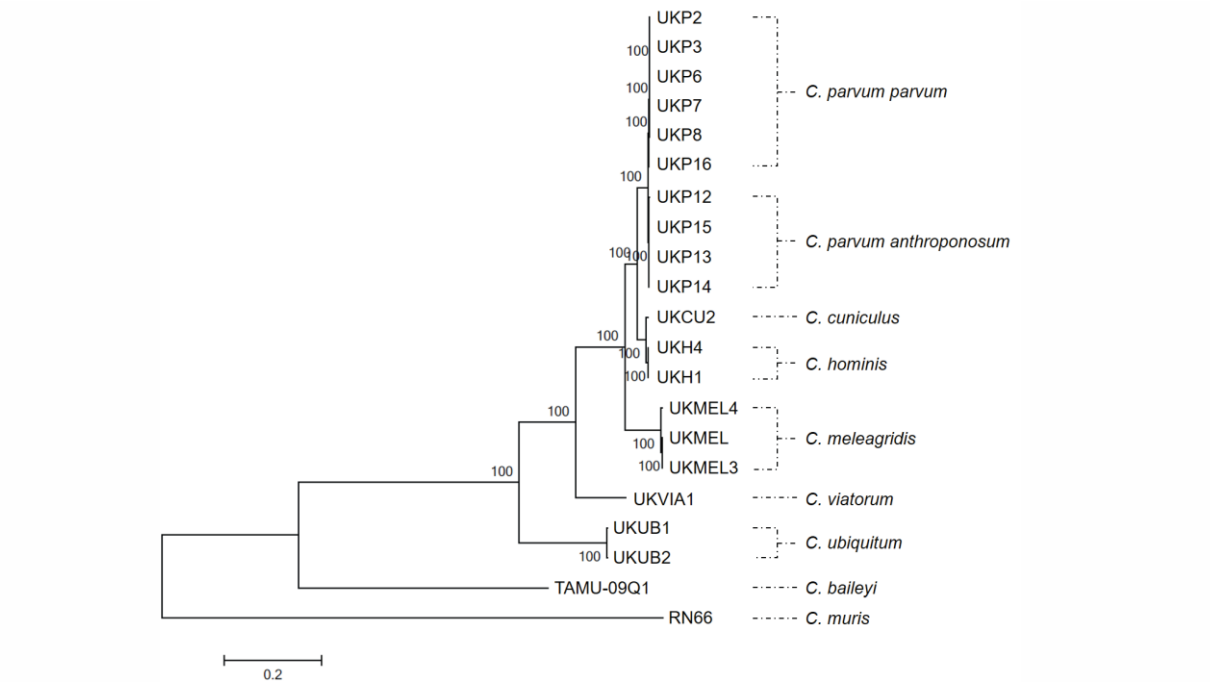
Supplementary Figure 1

Host ranges for human-infective *Cryptosporidium spp*. gp60 subtype families from GenBank-submitted gp60 sequences. Host ranges were determined for *C. hominis* gp60 subtypes Ia (N=327) and Ib (N=1752), *C. p. anthroponosum* IIc-a (N=111), and *C. p. parvum* subtypes IIa (N=843) and IId (N=377). Host types were characterised as equine, human, marsupial, mollusc, rodent, ruminant, primate, and other.



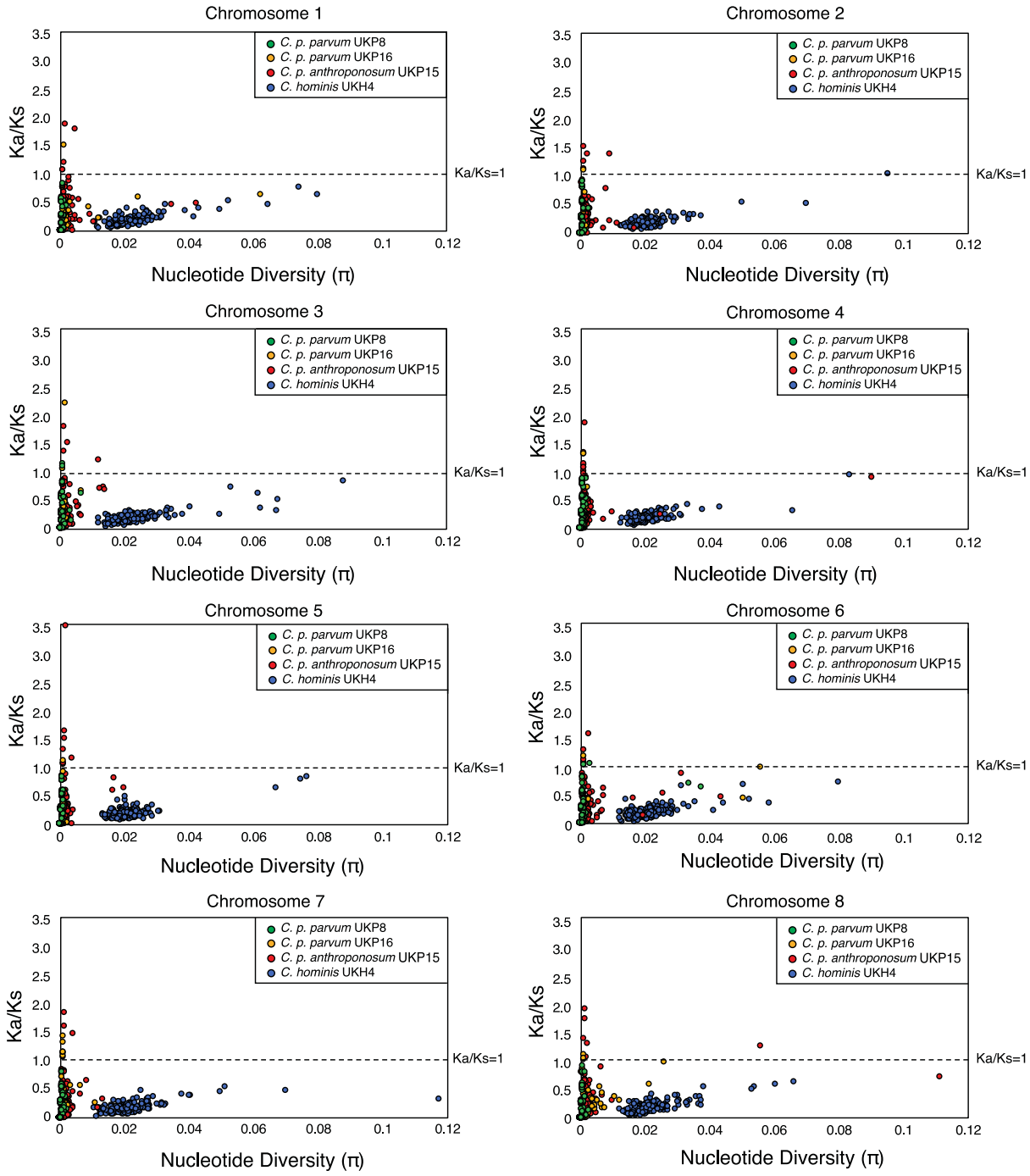
Supplementary Figure 2

Concatenated phylogeny of 21 human-infective *Cryptosporidium spp*. The maximum likelihood (ML) phylogeny based on a 153,421 bp alignment of 61 loci is shown. Included sequence targets exhibited neutral evolution between *C. p. parvum* UKP6 and *C. hominis* UKH4 (Ka/Ks 0.2-0.6, 93.0-98.0% nucleotide identities). Confidence values on the phylogeny reflect 2,000 bootstrap replications.[12]



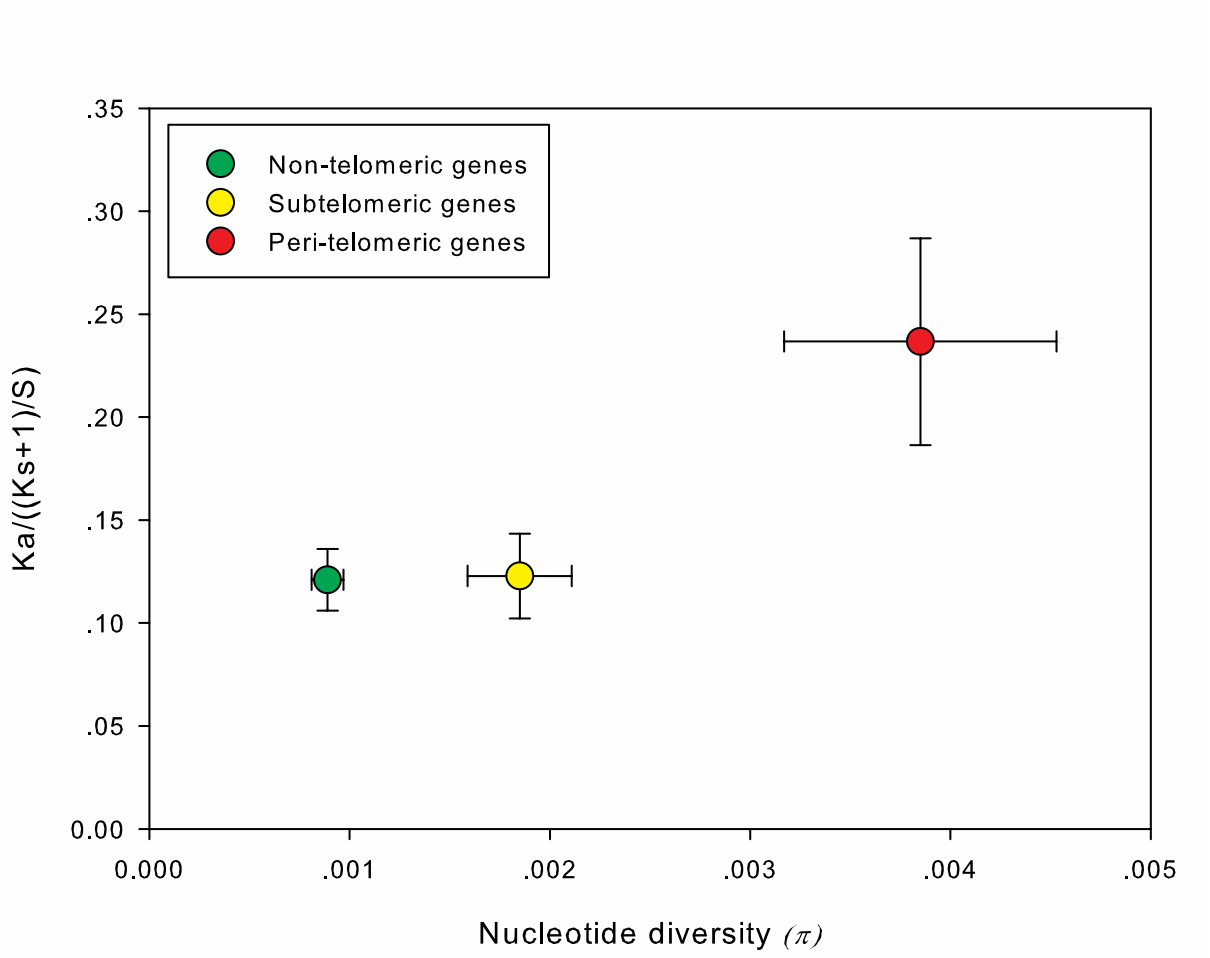
Supplementary Figure 3

Gene-by-gene signatures of selection (Ka/Ks) and nucleotide diversity (π) between human-infective *Cryptosporidium spp*. WGS across chromosomes 1-8. The nucleotide diversity is highest in *C. hominis* UKH4, whereas the signature of positive selection is most pronounced for *C. p. anthroponosum* UKP15.



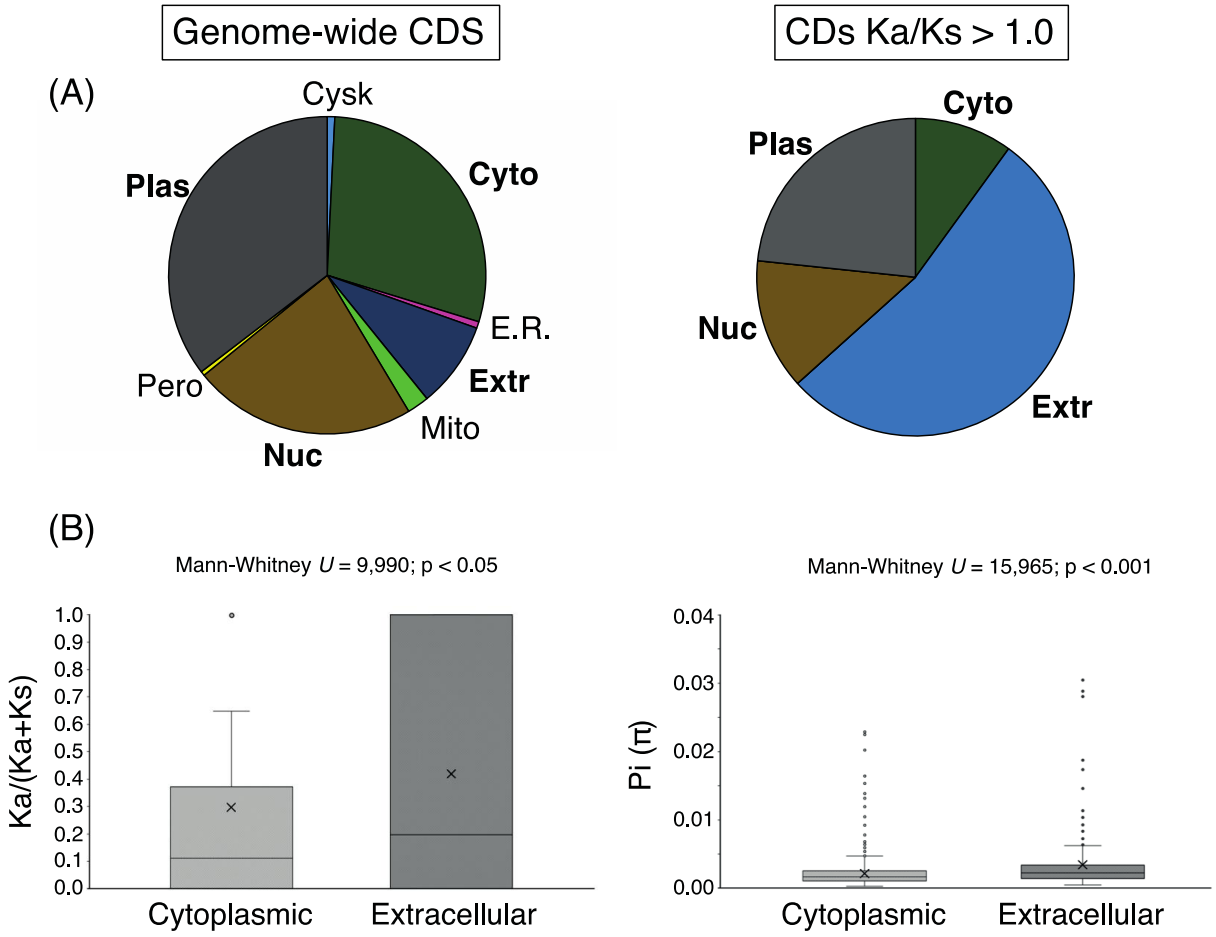
Supplementary Figure 4

Mean (±SE) nucleotide diversity (π) and signature of selection (Ka/((Ks+1)/S)) of genes in the non-telomeric (green, n=2827 CDSs), subtelomeric (yellow, n=326 CDSs) and peri-telomeric (red, n=312 CDSs) regions. Genes near the telomeres are the fastest evolving.



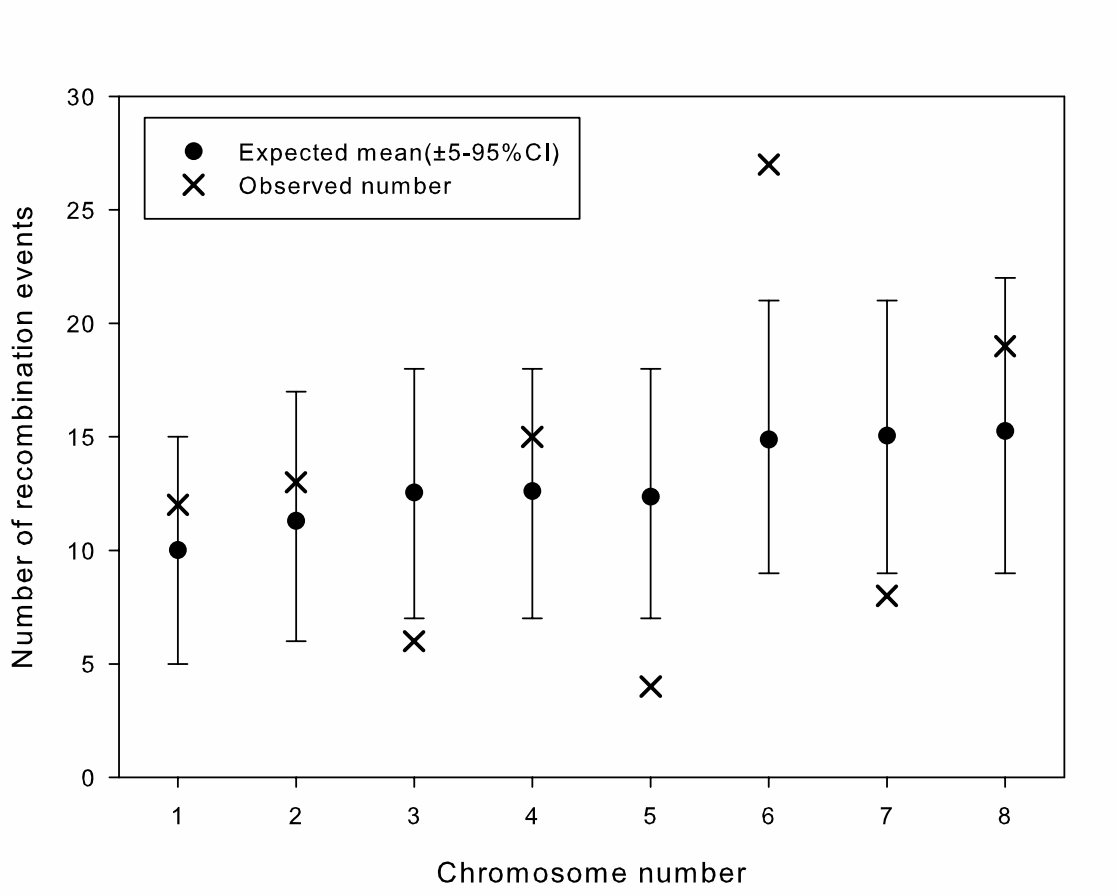
Supplementary Figure 5

(A) Predicted proportion of protein localization types for genome-wide CDSs and CDSs exhibiting significantly positive Ka/Ks values (>1.0), as compared between *C. p. parvum* UKP6 and *C. p. anthroponosum* UKP15. Protein localizations were categorised as cytoskeleton (Cysk), cytoplasm (Cyto), endoplasmic reticulum (E.R.), mitochondrion (Mito), nucleus (Nuc), peroxisome (Pero) and plasma membrane (Plas). (B) Comparative selective pressure (Ka/(Ka+Ks)) and nucleotide diversity (π) between CDSs annotated as having a cytoplasmic versus extracellular protein localization. Extracellular CDSs have a significantly faster rate of evolution (higher π) that is driven by positive selection (significantly higher Ka/(Ka+Ks) (two-tailed Mann-Whitney test n=3465 CDSs: Cytoplasmic n=1152 (Min=0.0000000, Median=0.0009709, Max=0.0375539), Extracellular n=333 (Min=0.0000000, Median=0.001311, Max=0.837771)). Exact p-value Mann-Whitney Ka/(Ka+Ks): p=0.0013. Exact p-value Mann-Whitney nucleotide diversity (π): p=1.233E-07.



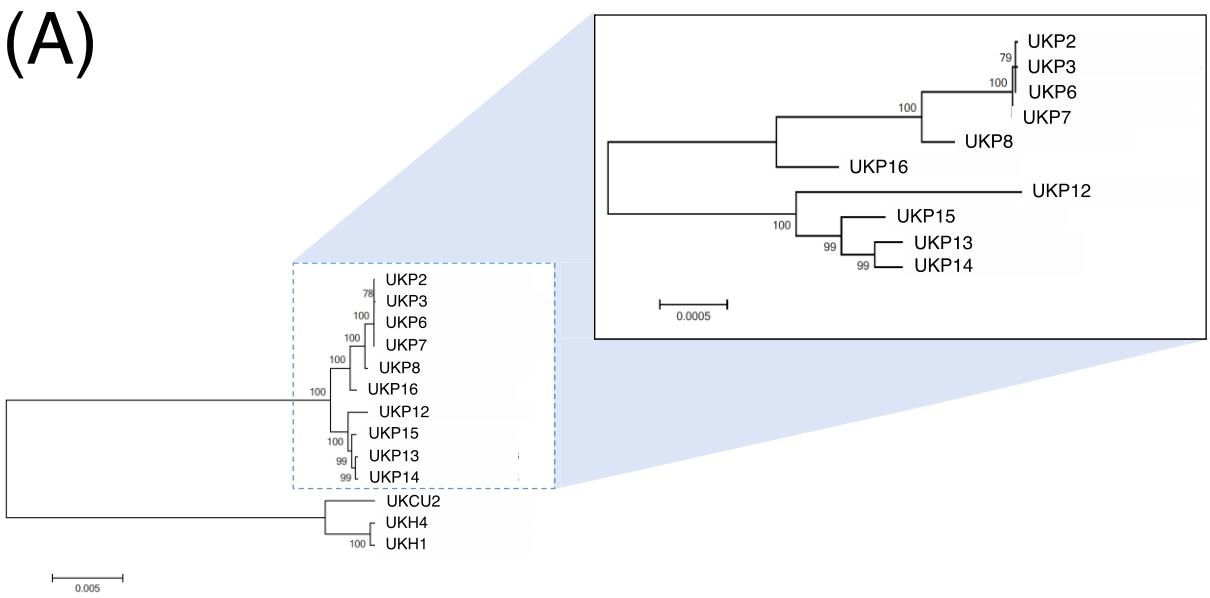
Supplementary Figure 6

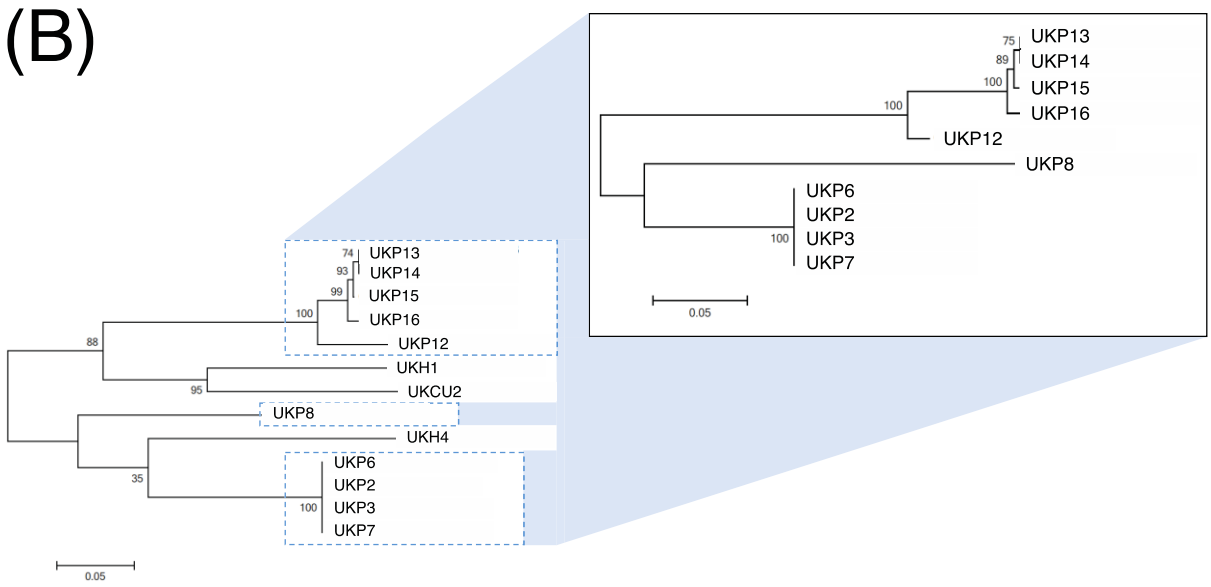
Mean and 5-95% confidence intervals of the expected number of recombination events per chromosome (based on chromosome size expressed as nucleotides) compared to observed number of recombination events in the RDP4 analysis (see Supplementary Table 2). The number of recombination events (n=104) are not homogeneously distributed across chromosomes, and chromosome 6 shows a significantly elevated number of events.



Supplementary Figure 7

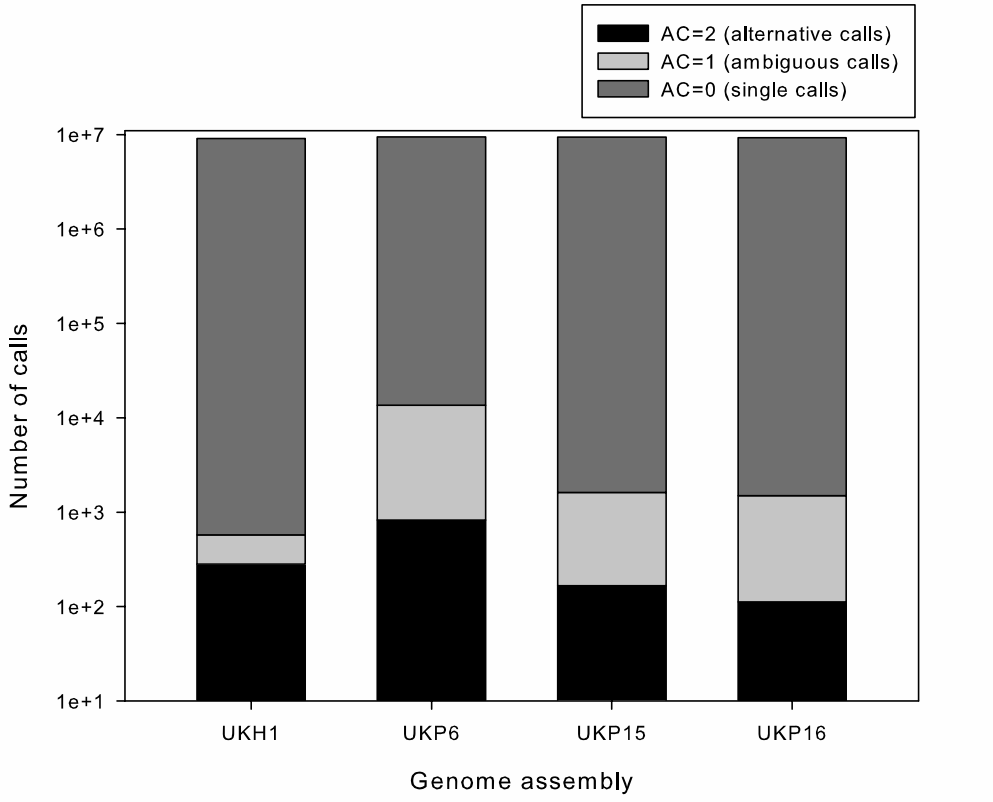
Incongruence between concatenated (A) and GP60-based (B) phylogenies of WGS used in this study. Zoomed sections illustrate phylogenies constructed using the same sequence alignments, but including only *C. parvum* WGS. This illustrates that the taxonomic relationships of the isolates based on the commonly used GP60 locus differs from that obtained by WGS, and that the GP60 locus alone cannot effectively resolve the evolutionary relationships between species. Trees were generated using the automated ClustalW alignment algorithm and Maximum Likelihood phylogeny builder, using 1000 bootstrap replications, in Mega 7.0.[12]





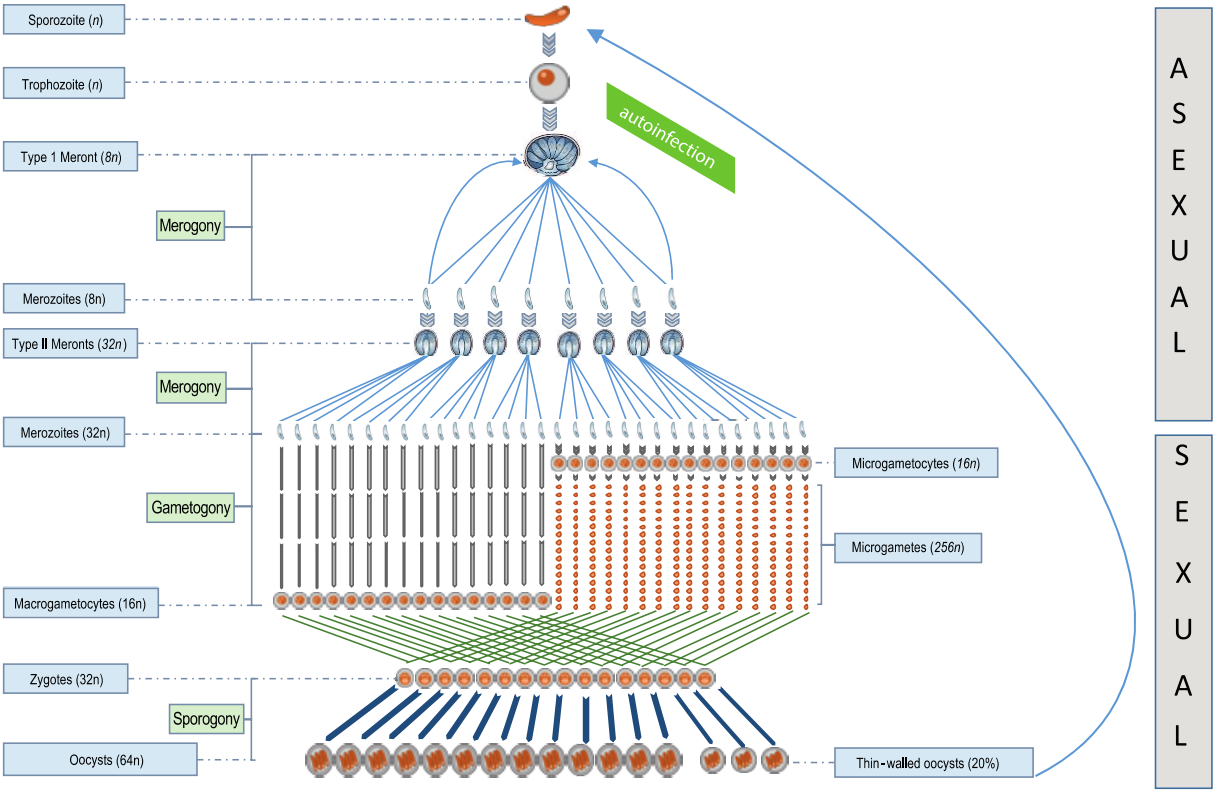
Supplementary Figure 8

Stacked bar graph of the number of calls of bases from the reads of the four isolates that were studied in the genetic introgression analysis (UKH1, UKP6, UKP15 and UKP16). Note that the Y-axis is log_{10}-transformed, and that the vast majority (>99.85%) of the calls are single bases (AC=0), which gives confidence that each of these four samples represent a single isolate. AC=0 represent “single called” bases for which there is no evidence of alternative calls. AC=1 indicates an ambiguous call, and AC=2 indicates a true alternative call. Such ambiguous and alternative calls are evidence of polymorphisms, which for this haploid species suggests either: (1) contamination from e.g. mixed infections, (2) polymorphisms arising due to novel mutations in the genome of parasite population accumulated whilst in the host, or (3) sequencing errors. For all four isolates examined, the fast majority of bases (>99.85%) were reliable assessed as “single calls” (i.e. AC=0). The UKP6 isolate had 0.134% of its bases called ambiguously (AC=1), and 0.009% bases called with an alternative base (AC=2). This represents a very small fraction of the genome in total, which gives confidence that each of these four samples represent a single isolate.



Supplementary Figure 9

Illustration of a *Cryptosporidium* generation[13,14]
Schematic illustrates the required rounds of DNA replication to complete the *Cryptosporidium* life-cycle. Oocysts in the environment contain four haploid sporozoites which are released from thick-walled oocysts in the host after ingestion. Each sporozoite is infective, forming a trophozoite following infection and invasion of an intestinal epithelial cell. Three rounds of DNA replication – merogony – follow, forming a type 1 meront which releases 8 type I merozoites. Each type 1 merozoite is able to independently infect an additional epithelial cell and two further rounds of DNA replication follow to form a type 2 meront which releases 4 type II merozoites. Alternatively, type 1 merozoites can produce further type 1 meronts. Type 2 merozoites are able to undergo gametocytogenesis producing either single haploid macrogametocyte or (following four rounds of DNA replication) 16 haploid microgametes. The cycle is completed when fusion of a microgamete with a macrogametocyte produce a diploid zygote and the ensuing meiosis gives rise to oocysts with 4 haploid sporozoites. Oocysts are either thick-walled environmentally resistant forms or thin walled forms that lead to autoinfection. (n = one haploid genome. The proportions/numbers of parasites shown progressing through the life-cycle are approximated for illustrative purposes).



References

1. Hadfield, S. J. et al. Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples. *BMC Genom*. **16**, 650 (2015).

2. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452 (2009).

3. Puiu, D. et al. CryptoDB: the *Cryptosporidium* genome resource. *Nucleic Acids Res.* **32**, D329–D331 (2004).

4. Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98 (1999).

5. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).

6. Horton, P. et al. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* **35**, W585–W587 (2007).

7. Apweiler, R. et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **32**, D115–D119 (2004).

8. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, vev003 (2015).

9. Ward, B. J. & van Oosterhout, C. HYBRIDCHECK: software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. *Mol. Ecol. Resour.* **16**, 534–539 (2016).

10. Okhuysen, P. C. et al. Infectivity of a *Cryptosporidium parvum* isolate of cervine origin for healthy adults and interferon-γ knockout mice. *J. Infect. Dis.* **185**, 1320–1325 (2002).

11. Chappell, C. L. et al. *Cryptosporidium meleagridis*: infectivity in healthy adult volunteers. *Am. J. Trop. Med. Hyg.* **85**, 238–242 (2011).

12. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).

13. Kosek, M., Alcantara, C., Lima, A. A. M. & Guerrant, R. L. Cryptosporidiosis: an update. *Lancet Infect. Dis.* **1**, 262–269 (2001).

14. O'Hara, S. P. & Chen, X. M. The cell biology of *Cryptosporidium* infection. *Microb. Infect*. **13**, 721–730 (2011).