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[38] . 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 a 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 all peer-reviewed journal publications describing non-outbreak associated prevalence studies in humans, published between 2000-2017. A total of 7,977 PubMed (https://www.ncbi.nlm.nih.gov/pubmed) hits for the search term "*Cryptosporidium*" were obtained, and initially filtered to remove all pre-2000 publications (N = 2,555). Abstracts for the 5,422 hits that remained were individually reviewed to extract all human cryptosporidiosis prevalence studies (N = 1,010), and subsequently cleaned to remove all outbreak and seroprevalence studies, as well as studies where a full version of the publication could not be accessed. This resulted in a dataset of 693 peer-reviewed studies spanning 113 countries worldwide. In order to enhance global coverage, pre-2000 published human prevalence studies were included for countries where post-2000 data was not available, which added another 50 hits from 13 countries (Austria, Azerbaijan, Belarus, Bolivia, Burma,

Burundi, Congo, Costa Rica, El Salvador, Honduras, Liberia, Trinidad and Tobago, and Turkmenistan). The final dataset consisted of 743 peer-reviewed journal publications, and spanned a total of 126 (64.9%) of the 194 countries worldwide (SI References). The database was subsequently constructed using manually-extracted data from each of these studies, and included, where available: date of publication, date(s) of study, study group(s) (including all descriptive characteristics pertaining to age, pre-existing chronic conditions, disease manifestations, and high-risk exposures such as animal contact), location (including where possible country, city, and rural versus urban environment), overall prevalence, prevalence for individual study groups with pre-defined medical conditions (e.g. HIV-positive, transplant recipients), proportion of prevalence attributable to known species/genotypes, and the proportion of *C. hominis* and *C. parvum* cases caused by individual subtype families.
For the purpose of illustrating heat-mapped proportions of *C. hominis* versus *C. parvum* prevalence worldwide, the 743 peer-reviewed prevalence studies in the database were filtered to remove all publications that did not identify *Cryptosporidium* isolates to the species level, or that identified *Cryptosporidium* species other than *C. hominis* or *C. parvum*. This resulted in a dataset of 222 peer-reviewed publications covering 77 countries. In order to further illustrate proportional differences in prevalence between *C. parvum* subtype IIA and IIC, studies reporting *C. parvum* were further filtered to obtain those that provided gp60 sequence information for these two subtypes. This resulted in a dataset of 81 peer-reviewed publications, from 41 countries worldwide.

Empirical Data
Whole genome nucleotide sequence data for *C. meleagridis* UKMEL and *C. hominis* UKH1 were retrieved from the *Cryptosporidium* genetics database resource CryptoDB (www.cryptoDB.org)[41]. Three *C. hominis* WGS (UKH3, UKH4, and UKH5) and five *C. parvum* WGS (UKP2, UKP3, UKP6, UKP7, and UKP6) were previously published[2] and acquired through accession numbers PRJNA253834, PRJNA253838, PRJNA253839, PRJNA253836, PRJNA253840, PRJNA253846, PRJNA253847 and PRJNA253848, respectively. The remaining 11 *Cryptosporidium spp.* genome datasets were provided by primary investigators of a novel IMS/Illumina-based method for whole genome sequencing (WGS) of clinical isolates[42]. Basic specifications for whole genome sequences are outlined in Supplementary Table 1.

Concatenated Phylogenetic Analysis
Loci were selected for the concatenated phylogeny based on orthologous pairs of genes from the *C. parvum* UKP6 and *C. hominis* UKH4 genomes exhibiting a neutral degree of selective pressure (Ka/Ks: 0.3 – 0.6), a minimal degree of protein-level variation (93 - 96% amino acid identities), and an even genome-wide distribution (See Table S10). Initial selections were derived from a whole-genome Ka/Ks comparison dataset generated between *C. parvum* Iowa II and *C. hominis* TU502 (KaKs Calculator v. 2.0)[43]. Selection criteria were reconfirmed by locating genes in novel genome datasets belonging to the closest gp60 relatives of Iowa and TU502 (UKP6 and UKH4 respectively) by BLASTn searches of standalone genome databases (BioEdit v7.2.5)[44]. Coding sequences were subsequently confirmed (ORF Finder, NCBI; http://www.ncbi.nlm.nih.gov/projects/gorf/)[45], translated (In-Silico Sequence Conversion Tool; http://in-silico.net/tools/biology/sequence_conversion), aligned and compared (EMBOSS Stretcher EMBL-EBI)[46] and Ka/Ks recalculated[43]. Geographic location was ascertained through NCBI's chromosome annotation tool (NCBI Map Viewer)[47]. Orthologous coding sequences in the remaining whole genome *Cryptosporidium spp*. datasets were similarly identified by forward and backward BLASTing between standalone databases[44] of annotated protein coding sequences (EMBOSS GetOrf)[46].

Coding sequences were extracted from whole genome nucleotide FASTA files, translated into protein coding sequences (In-Silico Sequence Conversion Tool; http://in-silico.net/tools/biology/sequence_conversion) and concatenated manually into MEGA v. 7.0 (Molecular Evolutionary Genetics Analysis)[48]. Coding sequences associated with inter-contig ambiguities in any of the 20 whole genomes or that revealed >7.0% amino acid divergence in *C. hominis* or *C. parvum* datasets were excluded. Sequences were aligned using the ClustalW algorithm imbedded in MEGA with default values, and transformed into trees using the Maximum Likelihood (ML) method. Phylogenetic relationships between sequences were inferred using the Dayhoff substitution model, Nearest-Neighbour-Interchange (NNI) method, and 2,000 bootstrap replications.

Whole Genome Comparisons

Whole genome strains chosen for protein divergence characterization were included based on gp60 similarity to the zoonotic *C. parvum* Iowa II and anthroponotic *C. hominis* TU502 reference genomes (IIaA15G2R1 and IaA25R3 respectively), namely UKP6 (IIaA15G2R1) and UKH4 (IaA14R3). To explore the association between protein-level genome changes and host range specification, two further closely-related but phenotypically-diverse whole genome *C. parvum* strains were included; zoonotic strain UKP8 (IIdA22G1) and a common anthroponotic strain UKP15 (IIcA5G3a). These were specifically chosen based on equidistant gp60 nucleotide variability from *C. parvum* IIaA15G2R1, but different degrees of genomic-phylogenetic distance. Parallel comparisons of UKP15 and UKH4 against UKP6 made it possible to determine whether anthroponotic host range is associated with site-specific protein sequence divergence. The tertiary comparison between UKP8 and UKP6, both zoonotic, served as a positive control.

Whole genome datasets were annotated into protein coding sequences[46], uploaded as standalone protein databases[44] and cross-BLASTed locally using the BLOSUM62 substitution matrix and a cut-off expectation value of 1.0E-10. Tabulated results were visually scrutinized and all putatively divergent protein sequences (<90.0% amino acid identities) were manually extracted. Amino acid coding sequences were re-checked (ORF Finder), underwent functional characterization (UniPROT BLASTp; E-threshold < 1.0E-5; http://www.uniprot.org/blast/)[49], and protein localization (Wolf PSORT)[50]. Protein localization through Wolf PSORT was based on a *k*-nearest neighbour classifier algorithm, where queried sequences were assigned ranked predictions of localization based on numerical localization features in proteins with known annotations. Prediction categories include nuclear, cytoplasmic, mitochondrial, extracellular, plasma membrane, cytoskeleton, endoplasmic reticulum, peroxisomal and golgi apparatus. Protein sequences were then backwards BLASTed (BioEdit) to identify putative paralogs and classify intra-species protein families. Significant stretches of sequence inconsistencies (indels > 3,000bp) between whole genomes were identified by global nucleotide alignment (EMBOSS Stretcher) of individual contigs from de novo assemblies. Chromosome ends were additionally re-assembled (30,000bp extending from the 5'-AAACCT-3' telomeric repeats) manually by identifying overlapping contig sequence data, with the exception of the 5' end of chromosomes 1 and 7, and the 3' end of chromosome 8, as these lack sequence evidence of telomeric repeats[41]. The nature of divergence between orthologous protein sequences was further characterized from a causative perspective. The presence of nucleotide indels resulting in frameshift mutations were ascertained by parallel protein and nucleotide coding sequence alignments, which revealed instances where nucleotide changes had shifted the start/stop codon upstream or downstream, or altogether eradicated the open reading frame (EMBOSS Stretcher). Non-consensus amino acid positions due to single amino acid gaps (SAAGs) or polymorphisms (SAAPs) were assigned a non-discriminatory value of 1.0, so that proportions of protein

divergence attributed to either could be inferred. The direction of selective pressure between conserved orthologous coding sequences was evaluated through Ka/Ks calculation, using the CodeML phylogenetic analysis component of PAML (PAL2NAL)[51]. Ka/Ks values smaller or equal to unity (Ka/Ks≤1.0) were interpreted as sites under purifying selection, and Ka/Ks values above unity (Ka/Ks>1) were considered to be under positive selection. Sliding window Ka/Ks analyses, indel characterisations, and F_{st} calculations were all performed using DnaSP v. 5.10.1[52]. In addition, we used the package NaturalSelection.jl (https://github.com/BioJulia/NaturalSelection.jl) to calculate the distribution of Ka and Ks values in a simulated population assuming neutral evolution. In these simulations, an ancestral sequence of each gene was first generated based on the phylogenetic relationship between the observed sequences in our data. This ancestral sequence was then mutated to generate a population of sequences with the same nucleotide diversity as in our data. The mutations were simulated using the observed transition : transversion ratio. This procedure was repeated 1000 times, and the null distribution of Ka and Ks values was thus established. The observed Ka/Ks values were then compared to identify outlier loci with Ka/Ks values above the 95%CI of simulated values (indicating genes under positive selection).

Recombination Analysis

Whole genome recombination analyses were performed as an adjunct to findings from the concatenated phylogeny. The closer genomic-phylogenetic proximity of *C. parvum* gp60 strain IIcA5G3j (UKP16) to the reference IIa subtype (UKP6) strain, rather than the classic IIcA5G3a subtype strain (UKP15), warranted investigating the presence, location, and chronology of potential recombination events between these three whole genome sequences. *C. hominis* strain UKH1 was included as an outgroup, due to its superior sequence coverage and assembly quality compared to other *C. hominis* WGS. In order to ascertain that the called nucleotide bases were derived from a single genotype and not from multiple diverged genotypes present in a reads (e.g. due to mixed infections), we counted the AC values of all bases in the four isolates that were studied in the recombination analysis. Poor quality bases, adaptor sequences and reads less than 36 base pairs (bp) long were removed using Trimmomatic[53]. The reads were then mapped to the respective genome assemblies for each isolate with Bowtie2[54] using the "--sensitive-local" mapping parameters, and a 5 bp trim applied to each end. Pilon[55] was then run with default parameters, to fix the SNPs and indels only (i.e. the "--fix bases" option) and to output a VCF file of the sequence variants. Recombination signals were identified and described to include potential major and minor parentage and recombinant fragment breakpoints using RDP4[56]. Automated detection algorithms RDP, GENECONV, Bootscan, Maxchi, and Chimaera were run with default values, with RDP p-values of $<10^{-5}$ used to infer confidence of recombination signals. Highly significant recombination signals (p-values $<10^{-30}$) were confirmed by crosschecking sequence integrity between WGS of the same GP60 subtype.

Phylogenomic analysis

Sequence reads for 16 *Cryptosporidium* isolates (Supplementary Table 1) were aligned to the *C. parvum* Iowa reference genome[48] with XXX and SNPs identified using SAMtools[57] and BCFtools[58]. Low quality SNPs were filtered out based on quality scores and sequence coverage using BCFtools filter (--include '%QUAL >= 30 && ((FORMAT/DP) < 750) && ((FORMAT/DP) >= 3) && (AVG(FORMAT/DP) > 25) && (AVG(FORMAT/DP) < (3 * AVG(FORMAT/DP)))) and only biallelic SNPs were retained. For each sample, we created a pseudoreference with filtered biallelic SNPs inserted using GATK FastaAlternateReferenceMaker[59]. We investigated relationships between *C. parvum* isolates using principle component analysis. The filtered set of biallelic SNPs were pruned

based on linkage disequilibrium using 10 kb sliding windows to generate a set of un-linked sites. The set of pruned sites were then used for principle component analysis with SNPrelate[60]. We estimated maximum likelihood (ML) phylogenies in 50 SNP fixed windows across the genome using RAxML[61] with a GTRCAT substitution model and ascertainment bias correction enabled[62], using the full set of biallelic SNPs for all mapped isolates. Each phylogenetic tree was made ultrametric using the *chronopl* function in APE[63] and a consensus phylogeny generated with DensiTree 2[64]. Topology weighting 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*). Population genetics parameters were also estimated in 50 Kb sliding windows across the genome with a 10 Kb step size using scripts from the genomics_general repository (https://github.com/simonhmartin/genomics_general). Sites were filtered using the same thresholds as for the phylogenomic analysis. For the calculation of absolute divergence (d_{xy}), all sites meeting the quality thresholds were retained regardless of whether any polymorphism was observed and low quality sites masked.

Dating recombination events

We estimated the timing of recombination events identified with RDP4[56] between *C. p. parvum* UKP6, *C. p. parvum* UKP16, *C. p. anthroponosum* UKP15 (pink) and *C. hominis* UKH1 using HybridCheck[65], which dates recombination (introgression) events in terms of generation numbers. Divergence estimates were based on the quantitative presence of mutations in introgressed blocks between two sequences, taking into account a baseline mutation rate (μ), as well as different substitution models with distinct base frequency and mutation rate assumptions[65]. In this study, the HKY85 substitution model was selected, as it assumes unequal base frequencies between sequences, a relevant prerequisite for the considerably AT-rich *Cryptosporidium* genome. The default mutation rate of 10^{-8} per generation was retained, as it fell within the conservative estimated mutation accumulation rate observed between two outbreak WGS that were sampled seven days apart (Table S5). Although these strains cannot be directly linked epidemiologically, they both originated in the North of England and were sampled during a large *Cryptosporidium* outbreak affecting the region during 2012. The calculated SNP mutation accumulation rate is also comparable to estimates for other eukaryotic unicellular microorganisms[66,67].

The method used to assign units of time to each generation of oocysts was based on a back-extrapolation of "oocysts out" versus "oocysts in", taking into account challenge dose, duration, and total excretion from past infectivity studies, as well as the proportional relationship between offspring and parental oocysts (Figure S9). Important to calculating this proportion was initially an understanding of the sex ratio between gametocytes produced during gametogony, as an unequal rate directly influences the number of progeny that are produced. A single published source makes reference to this rate in *Cryptosporidium*, estimating the proportion of male gametocytes at 0.5[68], or a 1:1 ratio between males (microgametocytes) and females (macrogametocytes). Similar rates have been observed in a number of other intestinal protozoa, including Eimeria[69] and Toxoplasma[70]. Another important consideration was the number of oocysts that are excreted (thick-walled) versus recycled (thin-walled) within the host. This number has previously been estimated at around 80% and 20%, respectively[71].

In light of these biological numerical considerations, taking into account the number of nuclear divisions within an infectious cycle[72], a factor of 12 autoinfective sporozoite offspring per 1 parental sporozoite was used to estimate the rate of expansion of oocyst populations in vivo. Using this value in correlation with excretion data from past infectivity studies revealed a population expansion of 3-5 new generations, and an estimated life cycle

duration of 2-4 days, or 48-96 hours, per infection (Table S6). Although this estimate deviates significantly from the frequently-reported generation time of 12-14 hours for *Cryptosporidium*[73,74], these resources do not make reference to the source of this estimation, which makes it difficult to verify or investigate how these numbers were achieved. In addition, an early cell culture experiment of *Cryptosporidium* development in human and animal cell lines showed that the sexual forms (macro- and microgametes) did not develop until 48 hours after inoculation, and that oocysts only appeared after 72 hours[75]. These estimates nicely complement the 48 to 96-hour generation estimate calculated in this study, lending support to the use of this range to generate estimated divergence dates for human-infective *Cryptosporidium spp.* Note that we assumed a steady rate of transmission within host populations, which implies that we report the minimum estimates of divergence times, because oocysts may be dormant outside the host. When dormant outside the host, mutations may not accumulate as fast (because there is no cell division), and hence, the number of mutations observed may actually have accumulated over a longer period of time, rendering our reported dates minimum estimates.

Population Genetic Analyses

The population genetic analysis included a total of 467 gp60 sequences submitted to the NIH genetic sequence database (Genbank)[76], collected in 43 countries between 2000 and 2017. These sequences were obtained independent from the systematic review of *Cryptosporidium* prevalence studies, and exclusively utilized physical sequence data submitted to and available through Genbank. The sample consists of 361 *C. p. parvum* IIa and 106 *C. p. anthroponosum* IIc-a gp60 sequences. The GenBank references used in this analysis can be retrieved by performing an advanced search of '*Cryptosporidium parvum*' and 'gp60', 'gp15', 'gp40', and 'gp15/40'. Population genetic structure was visualised using Fluxus network using median joining setting[77]. Circle size illustrates the frequency of the corresponding gp60 gene, and the length of the branch corresponds to the number of nucleotide differences. Isolation-by-distance analysis was performed using a linear regression analysis of the genetic distance (Kxy) between isolates (as response variable), and geographic distance between the sampling locations (as the independent variable).

Heterogeneity across the genomes in the number of SNPs (nucleotide variation, π), indels, and recombination events was tested with Chi-square tests and binomial tests. These tests examined differences in between chromosomes, chromosomal regions, recombinant regions, and genes. Chromosomal regions were arbitrarily divided into peri-telomeres (10.0% of all nucleotides of a chromosome), subtelomeres (10.0% of nucleotides), and non-telomeric (80.0%), to describe locations of genes that were adjacent (peri-telomeric) or proximal (subtelomeric) to 5'-AAACCT-3' telomeric repeats identified at the ends of whole chromosome sequences. Genes were divided depending on whether they showed a signature of positive selection ($Ka/Ks>1$) or not ($Ka/Ks\leq1$), and on whether they encoded proteins with predicted extracellular or cytoplasmic localizations[50]. Differences in nucleotide substitution patterns (Ka/Ks and Ka/(Ka+Ks)), indels and recombination events between species and strains were analysed using Mann-Whitney test and ANOVAs. The relationship between genetic differentiation and geographic distance between pairwise samples (i.e. isolation-by-distance) was examined using a Regression analysis. All tests were conducted in R (3.0.2; R Core Team 2013)[78] 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]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).

[43]Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics *Bioinformatics* **8**, 77-80 (2010).

[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**, pp. 95-98 (1999).

[45]Wheeler, D. L. *et al*. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31**, 28-33 (2003).

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

[47]Sayers, E. W. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. **38**, D5-16 (2010).

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

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

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

[51]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).

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

[53]Bolger, A. M., Lohse, M., & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

[54]Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357 (2012).

[55]Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**, e112963 (2014).

[56]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).

[57]Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).

[58]Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993 (2011).

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

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

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

[62]Leaché, A. D. *et al*. Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. *Systematic Biology* **64**, 1032-1047 (2015).

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

[64]Bouckaert, R. R. & Heled, J. DensiTree 2: Seeing Trees Through the Forest. *Unpublished; University of Auckland* (2014).

[65]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**, 534-9 (2016).

[66]Lynch, M. *et al*. Genetic drift, selection and the evolution of the mutation rate. *Nature Review Genetics* **17**, 704–714 (2016).

[67]Beghain, J. *et al*. *Plasmodium* copy number variation scan: gene copy numbers evaluation in haploid genomes. Malaria Journal **15**, 206 (2016).

[68]West, S. A., Smith, T.G. & Read, A. F. Sex allocation and population structure in apicomplexan (protozoa) parasites. *Proc Biol Sci*. **267**, 257-63 (2000).

[69]Chauve, C. M., Reynaud, M. C. & Gounel, J. M. Description d'*Eimeria mulardi* N. sp. chez le canard mulard. Etude de la phase endogene de son cycle evolutif avec mise en evidence du developpement intranucleaire. *Parasite* **1**, 15-22 (1994).

[70]Omata, Y. *et al*. Isolation of coccidian enteroepithelial stages of *Toxoplasma gondii* from the intestinal mucosa of cats by Percoll density-gradient centrifugation. *Parasitol Res*. **83**, 574-7 (1997).

[71]Ridley, R. K. & Olsen RM. Rapid diagnosis of bovine cryptosporidiosis with a modified commercial acid-fast staining procedure. *J Vet Diagn Invest*. **3**, 182-3 (1991).

[72]Kosek, M., Alcantara, C., Lima, A. A. M. & Guerrant, R. L. Cryptosporidiosis: an update. *Lancet Infect Dis*.**1**, 262-9 (2001).

[73]Upton, S. J. "Basic Biology of *Cryptosporidium*." Division of Biology, KSU. *Kansas State University*. Web (2008).

[74]Fleming, R. "*Cryptosporidium*: Could It Be in Your Water?" Factsheet. *Ontario Ministry of Agriculture, Food, and Rural Affairs*. Web (2015).

[75]Current, W. L. & Haynes, T. B. Complete development of *Cryptosporidium* in cell culture. *Science* **224**, 603-5 (1984).

[76]Benson, D. A. *et al*. GenBank. *Nucleic Acids Res*. **41**, D36-42 (2013).

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

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

[79]Kosek, M., Alcantara, C., Lima, A. A. & Guerrant, R. L. Cryptosporidiosis: an update. *Lancet Infect. Dis*. **1**, 262-9 (2001).

[80]O'Hara, S. P. & Chen, X. M. The cell biology of *Cryptosporidium* infection. *Microbes Infect*. **13**, 721-30 (2011).

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 222 peer-reviewed publications (SI References). 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 81 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.