Comparative genomics of *Cryptosporidium parvum* reveals the emergence of an outbreak-associated population in Europe and its spread to the United States

Greta Bellinzona,[1] Tiago Nardi,[1] Michele Castelli,[1] Gherard Batisti Biffignandi,[1] Karim Adjou,[2] Martha Betson,[3] Yannick Blanchard,[4] Ioana Bujila,[5] Rachel Chalmers,[6,7] Rebecca Davidson,[8] Nicoletta D'Avino,[9] Tuulia Enbom,[10] Jacinto Gomes,[11] Gregory Karadjian,[2] Christian Klotz,[12] Emma Östlund,[13] Judith Plutzer,[14] Ruska Rimhanen-Finne,[15] Guy Robinson,[6,7] Anna Rosa Sannella,[16] Jacek Sroka,[17] Christen Rune Stensvold,[18] Karin Troell,[13] Paolo Vatta,[16] Barbora Zalewska,[19] Claudio Bandi,[20] Davide Sassera,[1,21] and Simone M. Cacciò[16]

[1]Department of Biology and Biotechnology, University of Pavia, 27100 Pavia, Italy; [2]UMR BIPAR, Anses, Laboratoire de Santé Animale, INRAE, École Nationale Vétérinaire d'Alfort, 94700 Maisons-Alfort, France; [3]Department of Comparative Biomedical Sciences, School of Veterinary Medicine, University of Surrey, Guildford GU2 7AL, United Kingdom; [4]Viral Genetics and Biosecurity Unit (GVB), French Agency for Food, Environmental and Occupational Health Safety (ANSES), 22440 Ploufragan, France; [5]Department of Microbiology, Public Health Agency of Sweden, SE-171 82 Solna, Sweden; [6]Cryptosporidium Reference Unit, Public Health Wales, Swansea SA2 8QA, United Kingdom; [7]Swansea Medical School, Swansea University, Swansea SA2 8PP, United Kingdom; [8]Norwegian Veterinary Institute, N-1431 Ås, Norway; [9]Istituto Zooprofilattico Sperimentale dell'Umbria e delle Marche, 06126 Perugia, Italy; [10]Animal Health Diagnostic Unit, Finnish Food Authority, FI-70210 Kuopio, Finland; [11]National Institute for Agricultural and Veterinary Research, 1300 Lisbon, Portugal; [12]Department of Infectious Diseases, Robert Koch Institute, 13353 Berlin, Germany; [13]Swedish Veterinary Agency, SE-751 89 Uppsala, Sweden; [14]National Institute for Public Education, Budapest, 1007, Hungary; [15]Finnish Institute for Health and Welfare, FI-00271 Helsinki, Finland; [16]Department of Infectious Diseases, Istituto Superiore di Sanità, 00161 Rome, Italy; [17]Department of Parasitology and Invasive Diseases, National Veterinary Research Institute, 24-100 Pulawy, Poland; [18]Statens Serum Institut, 2300 Copenhagen, Denmark; [19]Veterinary Research Institute, Department of Food and Feed Safety, 621 00 Brno, Czech Republic; [20]Department of Biosciences, University of Milan, 20133 Milan, Italy; [21]IRCCS Fondazione Policlinico San Matteo, 27100 Pavia, Italy

The zoonotic parasite *Cryptosporidium parvum* is a global cause of gastrointestinal disease in humans and ruminants. Sequence analysis of the highly polymorphic *gp60* gene enabled the classification of *C. parvum* isolates into multiple groups (e.g., IIa, IIc, Id) and a large number of subtypes. In Europe, subtype IIaA15G2R1 is largely predominant and has been associated with many water- and food-borne outbreaks. In this study, we generated new whole-genome sequence (WGS) data from 123 human- and ruminant-derived isolates collected in 13 European countries and included other available WGS data from Europe, Egypt, China, and the United States (n = 72) in the largest comparative genomics study to date. We applied rigorous filters to exclude mixed infections and analyzed a data set from 141 isolates from the zoonotic groups IIa (n = 119) and IId (n = 22). Based on 28,047 high-quality, biallelic genomic SNPs, we identified three distinct and strongly supported populations: Isolates from China (IId) and Egypt (IIa and IId) formed population 1; a minority of European isolates (IIa and IId) formed population 2; and the majority of European (IIa, including all IIaA15G2R1 isolates) and all isolates from the United States (IIa) clustered in population 3. Based on analyses of the population structure, population genetics, and recombination, we show that population 3 has recently emerged and expanded throughout Europe to then, possibly from the United Kingdom, reach the United States, where it also expanded. The reason(s) for the successful spread of population 3 remain elusive, although genes under selective pressure uniquely in this population were identified.

[Supplemental material is available for this article.]

Corresponding authors: davide.sassera@unipv.it, simone.caccio@iss.it

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.278830.123. Freely available online through the *Genome Research* Open Access option.

The genus *Cryptosporidium* (phylum Apicomplexa) currently comprises at least 46 species and more than 120 genotypes of uncertain taxonomic status (Innes et al. 2020; Ryan et al. 2021a; Tůmová et al. 2023). Although the parasite has a global distribution, cryptosporidiosis represents a high-burden disease in children living in low-income countries, where it is a leading cause of moderate-to-severe diarrhea (Kotloff et al. 2013) and is associated with long-term negative impacts on childhood growth and well-being (Khalil et al. 2018).

Most *Cryptosporidium* species and genotypes have a narrow host range, suggesting coevolution with their hosts (Ryan et al. 2021a). Indeed, calibrated phylogenies indicate that much of *Cryptosporidium*'s diversity originated in the Cretaceous, as was the case for most of the mammals (Garcia-R and Hayman 2016). The mechanisms underlying host adaptation in *Cryptosporidium* are still poorly understood. Several species are known to infect different hosts, including *Cryptosporidium parvum*, *Cryptosporidium felis*, *Cryptosporidium canis*, *Cryptosporidium cuniculus*, *Cryptosporidium ubiquitum*, *Cryptosporidium meleagridis*, and others (Zahedi and Ryan 2020).

With no effective drugs and no vaccine, control of cryptosporidiosis is heavily dependent on the prevention of infection, which has to be informed by a detailed understanding of the epidemiology, population structure, and transmission dynamics of these parasites (Bhalchandra et al. 2018; Chavez and White 2018). The epidemiology of human cryptosporidiosis is complex, with transmission occurring indirectly via contaminated food or water or directly via contact with infected animals or individuals (McKerr et al. 2018). Most human cases are caused by *Cryptosporidium hominis*, which is anthroponotic, or *C. parvum*, which is zoonotic (Feng et al. 2018). Animal reservoirs, in particular young ruminants, have an essential role in the spillover and spillback of *C. parvum* to humans (Guo et al. 2022).

The most commonly used method for genotyping *C. parvum* isolates is by sequence analysis of the hypervariable gene coding for a 60 kDa glycoprotein 60 (*gp60*), which allowed delineating multiple groups, with IIa, IIc, and IId being the most common (Feng et al. 2018). In Europe, many IIa subtypes have been identified in humans, and many circulate among animals. However, a few subtypes appear to predominate, particularly subtype IIaA15G2R1, which is also the most common subtype globally (Chalmers and Cacciò 2016). The reasons for this high prevalence are unknown.

Recent studies based on whole-genome sequence (WGS) comparisons have begun to explore the evolutionary genetics of *C. parvum* (Feng et al. 2017; Wang et al. 2022; Corsi et al. 2023). In the work of Corsi et al. (2023), analysis of 32 WGSs indicated a clear separation between European and non-European (Egypt and China) isolates and highlighted the occurrence of recombination events between parasite populations. Another work analyzed 101 WGSs and hypothesized the existence of two ancestral populations, represented by IId isolates from China and IIa isolates from Europe. The authors proposed that the IId and IIa populations recently became sympatric in Europe, and generated hybrid genomes through recombination, possibly influencing biological traits such as host preference (Wang et al. 2022).

In this study, we generated WGSs for 123 human- or ruminant-derived *C. parvum* isolates collected across Europe. We also retrieved publicly available WGS data of 72 isolates from Europe, Egypt, China, and the United States, including the isolate IOWA-ATCC, which was used as a reference genome (Hadfield et al. 2015; Troell et al. 2016; Feng et al. 2017; Nash et al. 2018; Baptista et al. 2022; Wang et al. 2022; Corsi et al. 2023). Based on the largest comparative study to date, our main aim was to understand the evolution of this important zoonotic pathogen in Europe and in the United States.

Results

Quality control and sample selection

We started from an initial collection of WGS from 194 *C. parvum* isolates (including 123 newly sequenced isolates and 72 isolates retrieved from public databases) (for list, see Supplemental Table S1). To obtain a robust foundation for reliable inferences, we performed a careful selection based on multiple criteria, including the level of contamination from nontarget organisms, mean read depth, multiplicity of infection, and genome assembly quality (for more details, see Methods section) (Supplemental Table S2). This stringent selection process yielded a final data set of 141 isolates (including 88 newly sequenced isolates, 52 publicly available isolates, and the reference IOWA-ATCC), which were used for downstream analyses.

Importantly, the final data set comprised isolates from four continents (Africa, Asia, Europe, and North America) and from the two major zoonotic *gp60* groups (IIa and IId). Detailed information regarding the data set composition is provided in Supplemental Table S1.

Genetic variability among isolates

By comparing the 140 isolates to the IOWA-ATCC reference genome, 45,663 single-nucleotide polymorphisms (SNPs) and 18,909 insertion/deletions (indels), were identified. We filtered the SNPs in order to include only high-quality, biallelic SNPs (see Methods), which reduced the number to 28,047. The SNP distribution was not random at the level of individual chromosomes (Supplemental Table S3), with statistically significant higher SNP density observed at Chromosomes 1 and 6 (P-value $< 10^{-5}$) (Supplemental Table S4) and with an enrichment in subtelomeric regions compared with internal regions of the chromosomes (Supplemental Fig. S3).

Phylogenetic analysis and population structure

To provide a preliminary overview of the global evolutionary relationships among *C. parvum* isolates, we inferred phylogeny based on a defined set of 195 orthologous genes, and included *C. hominis* as an outgroup (Supplemental Fig. S4). We observed that all isolates from Europe and the United States formed a highly supported, monophyletic clade (hereafter, the "Western" lineage), whereas isolates from China and Egypt appeared to have diverged earlier.

Next, to obtain a more detailed description of the relationships among the 141 *C. parvum* isolates, we inferred a maximum likelihood (ML) tree based on the concatenated set of 28,047 biallelic SNPs (Fig. 1), using the root determined in the tree based on orthologous genes. The large-scale topology was consistent with the orthologous genes-based phylogeny. We observed that the host species and *gp60* subtypes were "scattered" along the tree, the latter indicative of limited predictive power for the deep phyletic relationships within *C. parvum*. However, a partial correlation with the geographical origin was found, as isolates sampled from the same region/country, sampled from a single farm, or collected within a narrow temporal window formed evident subclusters (e.g., all but one of the Finnish isolates formed a monophyletic clade). All isolates from the United States formed a fully supported monophyletic clade that was nested within the European isolates and exhibited a sister group relationship with a clade of isolates from the United Kingdom. Even upon excluding human isolates, we found no discernible changes in the topology of the

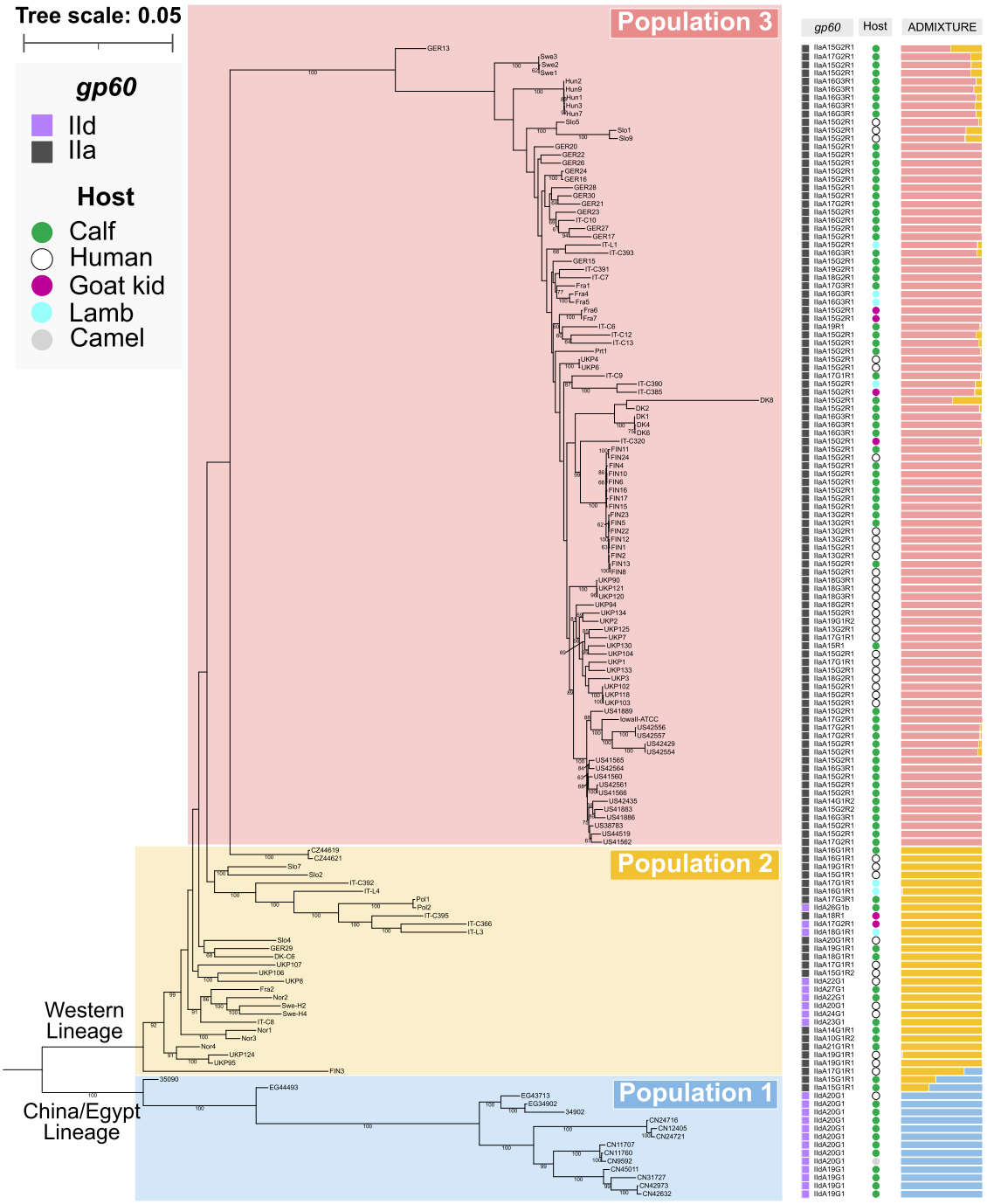

Figure 1. Maximum likelihood (ML) tree based on a set of 28,047 biallelic SNPs. Only bootstrap values greater than 60 are shown. Information about host, *gp60* subtype, and results of ADMIXTURE is mapped on the phylogeny.

phylogenetic tree or in the clustering patterns among animal species (Supplemental Fig. S5).

We then investigated population structure using ADMIXTURE and identified k = 3 as the most probable number of populations, in overall agreement with phylogeny (Fig. 1). Population 1 encompassed all the Chinese and Egyptian isolates (15); population 2 included a nonmonophyletic group of a minority of European isolates (28/109); and population 3 comprised the majority of the European isolates (81/109) and all those from the United States (17), which together formed a monophyletic clade.

Although the three populations were genetically very distinct (Fig. 1), admixed isolates were also evident, most notably the IIa isolates from Egypt and the European isolates FIN3, GER13, and DK8. A phylogenetic network showed several connections between isolates from different populations, suggestive of recombination events (Supplemental Fig. S6).

Recombination analyses

To infer recombination events that may have contributed to the formation of mosaic genomes, we conducted a comprehensive analysis for each chromosome and performed SNP-based phylogeny, pairwise divergence (Dxy), and SplitsTree. The results are detailed in Supplemental Figures S7–S21, and here below, we summarized the most salient results.

At Chromosome 1, phylogenetic analysis showed the IIa isolates from Egypt (35909 and EG4493) clustering with population 2 and not with population 1 (Supplemental Fig. S7A), in contrast with the topology based on all genomic SNPs (Fig. 1). An inspection of the SNP distribution revealed a mosaic pattern in which the IIa Egyptian isolates are either very similar to the IId isolates from Egypt and China (population 1) or to the IIa and IId isolates from Europe (population 2), as shown in Supplemental Figure S8. Indeed, in the region spanning position 755,934 to 768,672 (~15 kb), 260 SNPs are found in population 1 (including isolates 35909 and EG4493 from Egypt), whereas populations 2 and 3 have very limited genetic variability (Supplemental Fig. S8). Immediately after this block, the IIa isolates from Egypt are essentially identical to those from population 2 until position 823,729 (~55 kb), whereas the IId isolates differ from the reference genome by 560 SNPs in this region (Supplemental Fig. S9). Among the genes in the latter region, several encode for proteins with signal peptides (e.g., members of the SKSR and CpLSP gene families). This mosaic structure is confirmed by the results of a SplitsTree analysis (Supplemental Fig. S7B).

Furthermore, we observed that isolate FIN3, a human-derived isolate from Finland with a history of travel to the Canary Islands, occupied a position between populations 1 and 2 in the phylogenetic analysis and showed signs of admixture and loops connecting it to population 1 (Supplemental Fig. S10). Indeed, in a 50 kb region spanning from position 824,800 to 874,170, the isolate FIN3 shared 443 SNPs with the IId isolates from population 1 (Supplemental Fig. S10). This region contained several genes encoding for proteins with signal peptides (e.g., members of the SKSR and CpLSP gene families, cgd1_140, cgd1_150, cgd1_160). Therefore, FIN3 is a hybrid that resulted from a recombination event that involved population 1, as further supported by the results of a SplitsTree analysis (Supplemental Fig. S7B).

At Chromosome 2, in a 210 kb region spanning from position 384,000 to 594,000, the isolate DK8 (calf isolate from Denmark, belonging to population 3) shares 460 SNPs with isolates from population 2 and, less so, population 3 (Supplemental Fig. S12). This region contains more than 50 genes, among which the presence of a member of the secreted GGC gene family and of the insulinase-like peptidase family can be noted.

At Chromosome 4, a typical mosaic structure is evident in the first 8 kb adjacent to the 5′ telomere (Supplemental Figs. S14, S15). In this region, the IId isolates from China (except those from Hebei and Shanghai), all Egyptian isolates, and the European isolates from Norway (calf isolate Nor1), Slovenia (human isolates Slo1, Slo2, and Slo9), and Italy (lamb isolates IT-C392 and IT-L3) shared about 270 SNPs and differed from all other isolates of populations 2 and 3, which were essentially identical to the reference genome. Four genes are located in this subtelomeric region, all encoding for uncharacterized proteins.

At Chromosome 6, in the first 18 kb adjacent to the 5′ telomere, the isolate Ger-13 (calf isolate from Germany, belonging to population 3) shares about 200 SNPs with isolates from populations 1 and 2, whereas all other population 3 isolates are essentially invariant, being identical to the reference genome (Supplemental Fig. S18). Five genes are located in this subtelomeric region: four encoding for uncharacterized proteins and one for an IMP dehydrogenase/GMP reductase. Therefore, Ger-13 is a hybrid that resulted from a recombination event that involved population 1, as further supported by the results of the SplitsTree analysis (Supplemental Fig. S17B).

At Chromosome 8, in a 30 kb region spanning position 210,000 to 240,000, the isolate DK8 (calf isolate from Denmark, belonging to population 3) shares 107 SNPs with isolates from population 2 and 3, whereas the remaining isolates from population 3 are largely invariant (Supplemental Fig. S21). Five genes are located in this region and encode for one RNA helicase, two uncharacterized proteins, a protein with putative membrane domain, and a protein with AP2/ERF domain.

Genomic variability at the population level

We observed that out of the 28,047 SNPs identified in the entire data set of 140 genomes, only 1243 (4.4%) were shared by the three populations, whereas the majority was specific to a single population (Supplemental Fig. S22).

We calculated the pairwise SNP distances among the 141 isolates and observed the smallest distances in population 3 (range, three to 2528 SNPs; average, 892 SNPs). On the other hand, larger SNP distances were observed in population 2 (range, 56 to 3532 SNPs; average, 1930 SNPs) and in population 1 (range, 60 to 4437 SNPs; average, 2113 SNPs). Considering inter-population variation, we observed that population 1 exhibited the greatest genetic divergence from both populations 2 and 3, with an average of 5867 and 5241 SNPs, respectively, whereas population 2 and population 3 displayed a lower average SNP distance (3847 SNPs).

Furthermore, we observed 13 clusters of highly similar genomes (defined as differing by fewer than 50 SNPs) (Fig. 2) encompassing from two to eight isolates and all belonging to population 3. Notably, these clusters were formed by isolates from known outbreaks or from epidemiologically linked cases. As examples, the human-derived isolates UKP102, UKP103, and UKP118 were from an outbreak that occurred in March 2016, whereas isolates UKP90, UKP120, and UKP121 were from a distinct outbreak that occurred in April 2016. Another cluster of highly similar genomes was formed by five Hungarian calf isolates (Hun1, Hun2, Hun3, Hun7, and Hun9), collected at a single farm from Pest County at multiple but short time intervals (May–June 2020), thus representing clearly epidemiologically linked cases and a possible outbreak. Another cluster comprised three Swedish calf isolates (Swe1, Swe2, Swe6) collected at a single farm in the same year.

In all these cases, very high genomic similarity was observed (pairwise SNP distance < 50 SNPs), and the respective isolates formed monophyletic, highly supported clusters in the phylogenetic analysis (Fig. 1). Although the isolates under study were not specifically collected to address this question, our data suggest that a threshold of 50 SNPs may be used to identify highly related *C. parvum* strains, which may serve as an appropriate cutoff to confirm suspected outbreaks at the genomic level.

To further investigate relationships among isolates within each population, we undertook an identity-by-descent (IBD) analysis and constructed relatedness networks at 90% and 80% (i.e., at which the fraction of shared IBD is >90% or >80%). As shown in Supplemental Figure S23, networks were formed by isolates from outbreaks and from single farms, as expected, but also by isolates from specific geographic areas, a result compatible with geographically structured

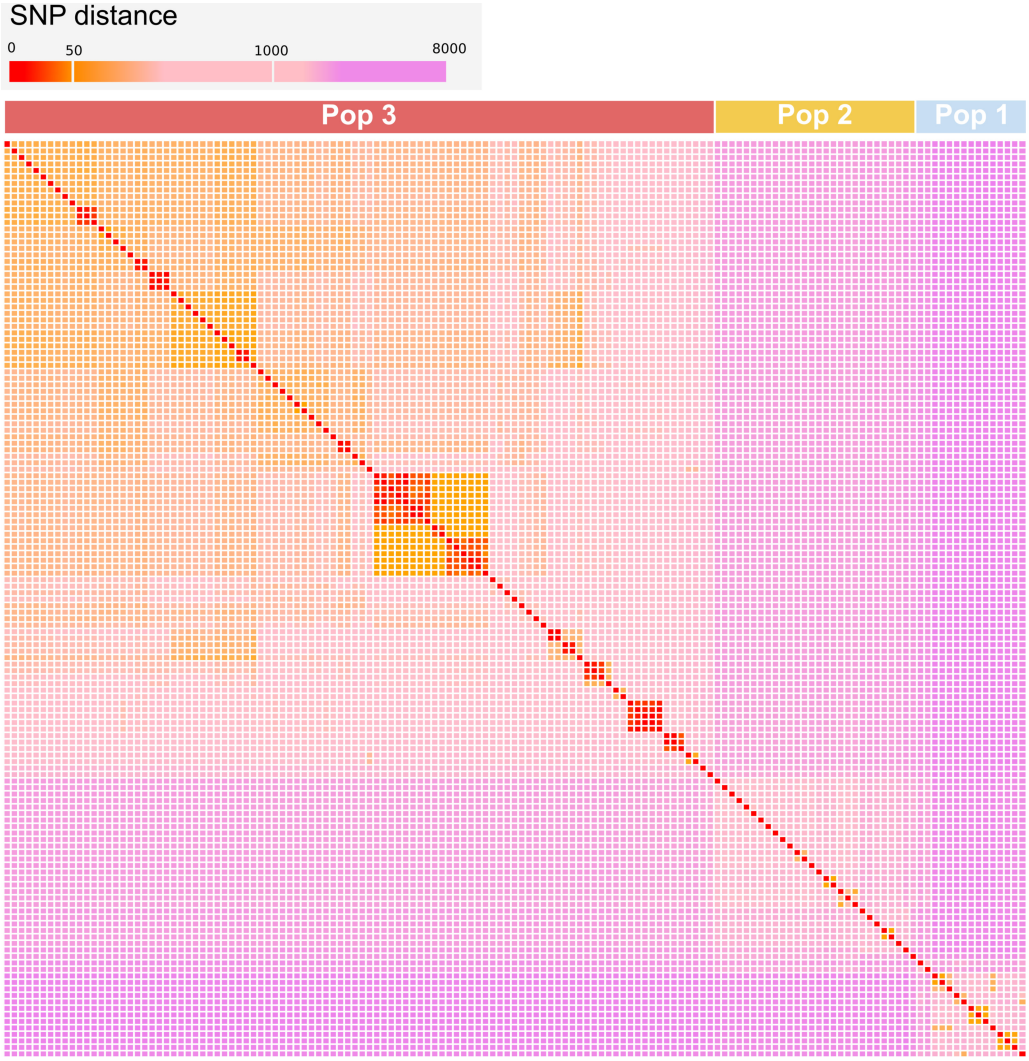

Figure 2. Heat map illustrating pairwise SNP distances among the 141 isolates analyzed. The order of the isolates reflects the position they occupy in the SNP-based phylogenetic tree. The color code is shown in the legend on the *top*.

populations. Notably, networks were observed within population 3 (Hungary, Finland, United States/United Kingdom, Germany/France) and population 1 (China, Egypt), but not for population 2.

Within-population genetic indices

To gain further insight into the genetic differentiation of the two populations in the "Western" lineage (i.e., populations 2 and 3), we calculated linkage disequilibrium (LD) decay, Tajima's *D* values, and nucleotide diversity (π).

We found that population 3 had a slower LD decay compared with that of population 2 (Fig. 3A), suggesting its more recent origin. We next observed that the distribution of Tajima's *D* values in population 3 was skewed toward negative values (Fig. 3B), indicating an excess of rare polymorphisms. This skewness was less pronounced in population 2 (Fig. 3B). Finally, we observed that population 3 exhibited lower nucleotide diversity compared with population 2 both at when analyzing the entire genome (respective means $\pi = 0.039$ and $\pi = 0.073$) and when analyzing single chromosomes (Fig. 3C).

These analyses are consistent with a recent origin and expansion of population 3, which can be explained by various factors, such as genetic drift (e.g., population bottlenecks) or selective sweeps.

We tested whether the presence of highly similar genomes in population 3 could have biased these results. We randomly selected a single isolate from each of the 13 clusters of highly related genomes and recalculated Tajima' *D* and π. LD decay analysis was not repeated, as this already involves use of equally sized subsets of random individuals from the two populations.

Although absolute changes in Tajima's *D* and π values were observed (Supplemental Fig. S24), the overall interpretation remained consistent with that obtained using the entire data set.

Genomic differences between population 2 and population 3

We investigated patterns of genetic variation between the two "Western" populations by first screening a set of 55 putative virulence factors involved in the host–parasite interplay (e.g., those encoding for mucin-like glycoproteins, thrombospondin-related

adhesive proteins, secreted MEDLE family proteins, insulinase-like proteases, and rhomboid-like proteases). We did not find any presence/absence pattern differentiating the two "Western" populations or when considering amino acid substitutions.

To investigate differences between population 3 and population 2 at the genome-wide level, we calculated the fixation index (F_{st}) in 1 kb windows (Fig. 4). This allowed the detection of genomic regions putatively under selection, and we inspected the genes present in such regions. By focusing on the top 1% of the total F_{st} values (i.e., applying a cutoff of 0.91), we identified 79 regions (Supplemental Table S5) and highlighted candidate genes under selection according to their function in Figure 4.

Next, we aimed to identify genes in which at least one nucleotide position exhibited significant ($P<0.05$) signs of selective pressure. Out of the 3385 annotated genes in the reference IOWA-ATCC, 228 appeared to be under selective pressure in population 3. To test whether these genes were under selective pressure exclusively in population 3, we extended our analysis to population 2. We found that 176 of 228 genes were under selection pressure exclusively in population 3 and not in population 2. Most of these genes (49/176) were annotated as "hypothetical proteins," whereas a few (eight of 176) were "putative secreted proteins" (Supplemental Table S6).

Notably, 16 proteins were identified in both analyses, that is, F_{st} statistics on genomic regions and selective pressure in single genes (Supplemental Table S7).

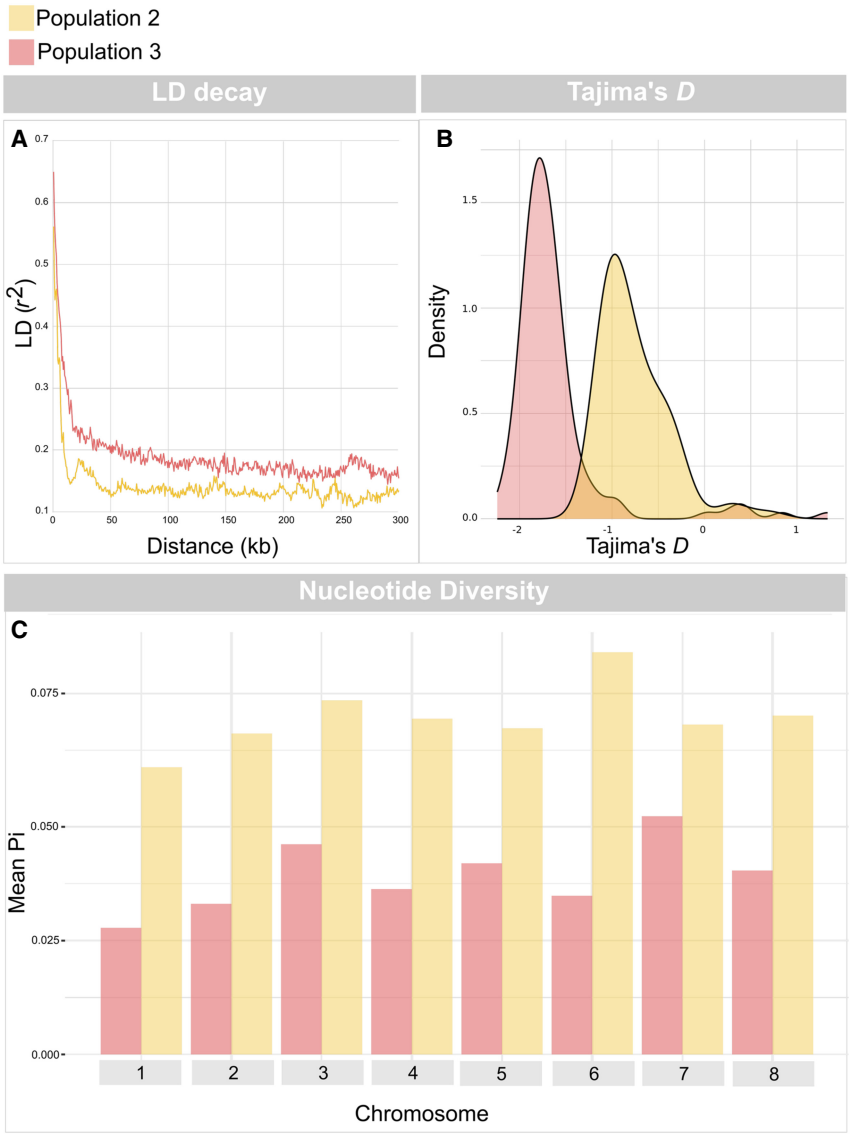

Figure 3. LD decay (*A*), distribution of Tajima's *D* values (*B*), and nucleotide diversity (π; *C*) in population 2 and population 3. Nucleotide diversity is shown with the mean π for each chromosome. Chromosome-specific results are provided in Supplemental Figure S8.

Discussion

C. parvum is the most prevalent zoonotic pathogen within the genus *Cryptosporidium* and is a global cause of diarrheal disease in humans and ruminants (Ryan et al. 2021b). This species is widespread in industrialized countries, including Europe (Cacciò and Chalmers 2016), but also in the Middle East (Hijjawi et al. 2022). Despite the recognized impact on human health and livestock production, no effective drugs or vaccines are available for controlling *C. parvum* infections. An urgent need for new control tools has been repeatedly underlined (Chavez and White 2018; Rahman et al. 2022; Khan and Witola 2023). Recent WGS studies have begun to provide insights into the genetics of *C. parvum*, proposing a role for recombination events in the evolution of this species, and have allowed identifying a number of genes under positive selection, potentially involved in host–parasite interactions (Wang et al. 2022; Corsi et al. 2023).

In this study, we conducted the most extensive comparative genomic analysis of *C. parvum* to date by generating WGS data from human- and ruminant-derived isolates collected in 13 European countries (n = 127). Additionally, publicly available WGS data from Europe, Egypt, China, and the United States (n = 71) were included (Supplemental Table S1; Hadfield et al. 2015; Troell et al. 2016; Feng et al. 2017; Nash et al. 2018; Wang et al. 2022; Corsi et al. 2023). We filtered the initial data set to obtain a thoroughly cleaned and curated data set that comprised 141 isolates (including the reference IOWA-ATCC genome), ensuring a robust foundation for reliable genomic analyses. Consistent with earlier findings (Baptista et al. 2022; Wang et al. 2022; Corsi et al. 2023), the overall genetic variability was modest, as just 28,047 biallelic high-quality SNPs were identified across the 141 genomes analyzed. Phylogenetic analyses using both orthologous genes and SNPs provided robust evidence for the presence of two distinct lineages (Fig. 1). One lineage (China/Egypt lineage) consisted of all Chinese (IId) and Egyptian (IIa and IId) isolates and was very distinct from the second lineage, in which all European (IIa and

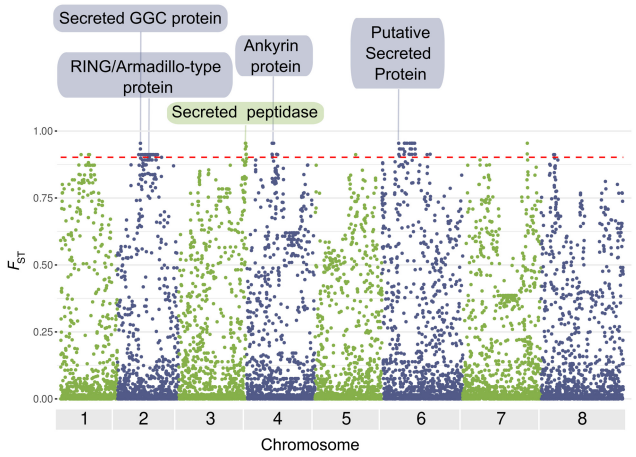

Figure 4. Manhattan plot of genome-wide Wright's F_{st} values, calculated in genomic regions of 1 kb, comparing population 2 and population 3. F_{st} values are shown on the *y*-axis and genomic positions on the *x*-axis. The dotted red line represents the cutoff value for the top 1%, equal to 0.91. Genes overlapping with the BUSTED analysis and with a function potentially associated with virulence or host–pathogen interactions are highlighted.

IId) and United States (IIa) isolates are grouped ("Western" lineage) (Fig. 1).

Considering that recombination has been, and still is, a fundamental driver of the genomic evolution of *C. parvum* (Wang et al. 2022; Corsi et al. 2023) and other *Cryptosporidium* species (Nader et al. 2019; Huang et al. 2023), we focused on tracing these events. We described two clear events occurring at Chromosomes 1 (Supplemental Fig. S10) and 4 (Supplemental Fig. S15), involving isolates from different populations and hosts, with representatives of population 1 being the putative minor parents (i.e., the donors). The very same event on Chromosome 4 has been already described on a more limited data set (Corsi et al. 2023), whereas here we provide extensive support for a recombination event on Chromosome 1 differing from that reported by Corsi et al. (2023). Additional evidence for the existence of mosaic genomes was obtained from network and admixture analyses of single chromosomes (Supplemental Figs. S7–S21). We observed SNP distribution patterns compatible with genetic exchanges, although the events could not be precisely reconstructed.

A deeper focus on the population structure showed that the "Western" lineage is subdivided into two groups, namely, population 3, a monophyletic group formed by all the United States and most of the European isolates, and population 2, a paraphyletic group that includes the remaining European isolates (Fig. 1). We propose that population 2 was the ancestral and more heterogeneous European population (including both IIa and IId isolates) and that population 3 (including only IIa isolates) has evolved more recently from it. Our hypothesis is strongly supported by the overall lower genetic diversity (π), negative Tajima's *D* values, and slowed decay in LD in population 3 compared with population 2 (Fig. 3).

Our reconstructions indicate that population 2, which includes both IIa and IId groups, is ancestral in Europe. Considering the relatively limited admixture herein evidenced with non-European isolates (Fig. 1), we hypothesize that the coexistence of the IIa and IId lineages in Europe could date back to ancient introductions of *C. parvum* from the Middle East, which is indeed one of the first areas in which livestock breeding originated (Beja-Pereira et al. 2006; Chessa et al. 2009). This is consistent with previous reconstructions based on the greater diversity of IId subtypes in Asia (Wang et al. 2014) and with the coexistence of IIa and IId in Egypt and several other Middle Eastern countries as well (Hijjawi et al. 2022).

A deeper focus on population 3 showed that the U.S. isolates from nine different states form a monophyletic clade, indicating a single event of introduction from Europe, likely from the United Kingdom (Fig. 1), and a subsequent expansion in the country. Historical data (Bowling 1942; Ficek 2019) and studies investigating the ancestry of New World cattle (McTavish et al. 2013; Delsol et al. 2023) suggested that this event should be relatively recent. Indeed, most of the import of livestock, particularly cattle, into the Americas occurred from the seventeenth to the nineteenth century by Portuguese and Spanish colonists and during the Victorian Age by British settlers (McTavish et al. 2013; Ficek 2019).

Our genomic results are consistent with *gp60* molecular typing data that identified only IIa subtypes in U.S. isolates (Jann et al. 2022). A parallel could be drawn to the recent emergence and rapid spread of the *C. hominis* IfA12G1R5 subtype in Europe, Australia, and the United States (Braima et al. 2019; Huang et al. 2023; Peake et al. 2023). In this case, however, the emergent lineage originated from successional recombination events involving North American, East African, and European populations (Huang et al. 2023), whereas in the case of the *C. parvum* population 3, the data support a single introduction in the country.

Moreover, we found that the clusters of isolates showing high genome similarity (<50 SNPs) all belonged to population 3 (Fig. 2), including those from known outbreaks. Thus, we investigated at genome-wide level the hypothesis of a selective advantage in population 3, which may explain its higher prevalence and association with water- and food-borne outbreaks. Although admittedly speculative, the most interesting result comes from a combination of population statistics and phylogeny-based statistical tests on gene sequences, which allowed identifying 16 candidate proteins under positive selection in population 3 alone. The candidates include genes encoding for secreted proteins, such as ankyrin repeat-containing proteins, which have been shown in *Toxoplasma* to be involved in cell invasion (Long et al. 2017), and a RING/Armadillo-type fold domain–containing protein that in *Plasmodium falciparum* mediates the motility of the parasite, essential for fertilization and transmission (Straschil et al. 2010). Although the exact functions of these genes and the biological implications of our observations require further investigation, their identification opens avenues for understanding the mechanisms underlying the potential selective advantage. Other nonmutually exclusive explanations should be explored, including variation in the copy number of genes encoding virulence factors (Xu et al. 2019), their differential expression, or a higher capacity to withstand standard water treatments and persist longer in the environment while maintaining infectivity.

This study has several limitations, particularly in relation to the uneven representation of parasite isolates in terms of hosts and geographic origin. Moreover, isolates from Europe were collected over a few decades, and some were known to be related (e.g., from outbreaks). Finally, we compared isolates collected from areas of the world where the epidemiology of cryptosporidiosis differs widely. Therefore, the results cannot be generalized but rather represent a testable hypothesis that must be confirmed by enlarging the data set to include additional human isolates and,

more generally, isolates from underrepresented areas of the world (e.g., South America, Middle East).

Nevertheless, our findings allow us to propose a scenario for the evolution of *C. parvum* in Europe, highlighting the presence of two sympatric populations, one of which recently expanded to become predominant in young ruminants and humans. We further show that this population comprises all isolates of the virulent and hypertransmissible IIaA15G2R1 subtype and all outbreak strains, suggesting a selective advantage, and has spread into the United States, likely from the United Kingdom.

Methods

Parasite isolates

The information available for the 195 *C. parvum* isolates from humans and ruminants included in this study is provided in Supplemental Table S1. The data set comprised 123 isolates sequenced in the present study, 71 isolates from previous studies (Hadfield et al. 2015; Troell et al. 2016; Feng et al. 2017; Nash et al. 2018; Wang et al. 2022; Corsi et al. 2023), and the recently assembled IOWA-ATCC genome (Baptista et al. 2022), which was used as a reference genome.

Oocyst purification, DNA processing, and sequencing

An aliquot of the 123 fecal isolates was used to extract genomic DNA and to identify the species and the *gp60* subtype, using previously published protocols (Alves et al. 2003; Ryan et al. 2003). The procedures for DNA purification and extraction are detailed by Corsi et al. (2023). In short, oocysts were purified from fecal specimens by immunomagnetic separation, treated with bleach, and used for genomic DNA extraction. Genomic DNA was subjected to whole-genome amplification (WGA) using the REPLI-g midikit (Qiagen), according to the manufacturers' instructions.

For high-throughput sequencing experiments, ~1 µg of purified WGA product per sample was used to generate Illumina Nextera XT 2 × 150 bp paired-end libraries, which were sequenced on an Illumina NovaSeq 6000 SP platform. Library preparation and sequencing were performed at the Institut du Cerveau (ICM) in Paris, France.

Data filtering and SNP calling

Raw reads of the 194 isolates were quality-checked and then preprocessed to remove low-quality bases and adapter sequences using Trimmomatic v.0.36 (Bolger et al. 2014), with default parameters. A series of sequential steps were then applied to select isolates and SNPs according to multiple criteria (Supplemental Fig. S1; Supplemental Table S1).

The presence of *Cryptosporidium* spp. sequences was verified using MetaPhlAn v. 3.0.13 (Beghini et al. 2021) and phyloFlash v. 3.4 (Gruber-Vodicka et al. 2020). Only isolates showing the presence of *Cryptosporidium* spp. were retained for further analyses. Among these, MetaPhlAn identified the presence of *C. hominis* sequences in two isolates, accounting for <5% of the total reads in one case and <2% in the other. These samples were excluded owing to significant presence of other contaminants according to both MetaPhlAn and phyloFlash.

The *C. parvum* IOWA-ATCC (Baptista et al. 2022) was used as a reference genome to map the filtered reads of each sample with Bowtie 2 v.2.5.0 (Langmead and Salzberg 2012) with default settings. PCR duplicates were then marked using Picard MarkDuplicates v. 2.25.4 (https://broadinstitute.github.io/picard/). Variant calling (SNPs and indels) was performed using the GATK's HaplotypeCaller v. 4.2.2.0 (DePristo et al. 2011; Van der Auwera and O'Connor 2020) with default parameters and the option -ERC GVCF. SNPs were removed if quality depth was less than 2.0, Fisher strand was greater than 60.0, mapping quality was less than 30.0, mapping quality rank-sum test was less than −12.5, read position rank-sum test was less than −8.0, and strand odds ratio was greater than 3.0.

The read depth and the number of missing sites were calculated for each isolates using VCFtools (Danecek et al. 2011), and isolates with a mean read depth <20× were discarded. The GVCFs were then imported into a GATK GenomicsDB using the function GenomicsDBImport, and a combined VCF was created using the GATK GenotypeGVCFs function.

To maximize the quality, SNPs were further filtered using BCFtools (Danecek et al. 2021) based on the following criteria: biallelic SNPs, quality score greater than 30, allele depth greater than 20, minor allele frequency greater than 0.005, and missing ratio less than 0.5.

The moimix R package (https://github.com/bahlolab/moimix) was then used to estimate multiplicity of infection. The FWS statistic, a type of fixation index to assess the within-host genetic differentiation, was calculated on the filtered SNPs. In pure isolates with haploid genomes, the FWS is expected to approach unity. Isolates with a FWS < 0.95 were excluded, as they were likely to represent multiple infections (Manske et al. 2012). Examples of infections with estimated multiplicity of infection equal to one or greater than one are presented in Supplemental Figure S2.

Cleaned mapped reads were assembled using Unicycler v.0.5 (Wick et al. 2017) with the --linear_seqs 8 option, which accounts for the presence of eight linear chromosomes in the reference assembly. Isolates with a genome size <8 Mb (the size of the reference genome is 9.1 Mb) were discarded, thus leading to the final data set (Supplemental Table S1).

Pairwise SNP distances (i.e., the number of SNPs found in all possible pairs in the sample population) were calculated using snp-dists v0.8.2 (https://github.com/tseemann/snp-dists) and visualized using the R package heatmap.2. The number of SNPs in non-overlapping windows of 1 kb across each chromosome was counted using VCFtools (--SNPdensity) (Danecek et al. 2011), and visualized using the R package ggplot2 (Wickham 2016).

To compare the SNP density between each chromosome, the pairwise comparisons for proportions test implemented in R (R Core Team 2021) was used, and the probability (*P*) values were adjusted using the Bonferroni correction.

Phylogenetic and population structure analyses

To ensure proper rooting of the tree inferred from genomic SNPs, we first generated a tree based on orthologous genes from the 141 *C. parvum* isolates of the final curated data set and used *C. hominis* TU502 (GCA_001593465.1) as outgroup. The gene sequences of the *C. hominis* isolate and of the reference genome *C. parvum* IOWA-ATCC, were downloaded from CryptoDB (Puiu et al. 2004). The AUGUSTUS algorithm (Stanke et al. 2006) was locally trained on the *C. parvum* IOWA-ATCC genome and then used to predict coding sequences for the remaining 140 isolates. A set of 195 genes, which have been used previously for phylogenomic analyses of Apicomplexan (Mathur et al. 2021), was searched using BLASTP on the orthogroups identified by OrthoFinder v2.5.4 (Emms and Kelly 2019) in our data set. Of these, orthologs of 179 genes were identified. Each ortholog was aligned with MUSCLE 5.1 (Edgar 2004) and concatenated. A ML tree was inferred on the concatenated alignment according to the model indicated by ModelTest-NG (HKY + F + I, BIC criteria) (Darriba et al. 2020) with RAxML v.8.2.12 (Stamatakis 2014), with 100 bootstrap pseudoreplicates.

Next, a concatenated set of genomic SNPs was created by converting the VCF into a FASTA file (https://github.com/edgardomortiz/vcf2phylip). A ML phylogenetic tree was inferred with RAxML v.8.2.12 (Stamatakis 2014) using the GTR + G model, as indicated by ModelTest-NG (Darriba et al. 2020), with ascertainment bias correction and 100 bootstrap pseudoreplicates. The same procedure was applied separately on the SNPs located into each of the eight chromosomes to obtain individual chromosome phylogenies.

Population structure analysis was performed with ADMIXTURE v1.3.0 (Pritchard et al. 2000), with the number of populations tested (K) ranging from one to 12. Phylogenetic networks were generated by using the Neighbor-Net algorithm implemented in SplitsTree v.5 (Huson and Bryant 2006).

Pairwise IBD was calculated using a hidden Markov model (Schaffner et al. 2018), and relatedness networks were generated using the R package igraph (Csardi and Nepusz 2006).

Recombination analyses

The sequence of each chromosome was reconstructed for each isolate by editing the reference IOWA-ATCC sequences according to the corresponding filtered SNPs using the GATK's FastaAlternateReferenceMaker function (Van der Auwera and O'Connor 2020). Then, multiple sequence alignments of each chromosome were analyzed by the Recombination Detection Program software, version 5 (RDP5) (Martin et al. 2015) using five algorithms (RDP, Geneconv, Bootscan, MaxChi, and Chimæra) implemented in this software. Only events supported by at least three algorithms and with a *P*-value cutoff of 10^{-5} were considered significant.

Population genetic analyses

Tajima's *D* values were calculated using snpR (Hemstrom and Jones 2023) in nonoverlapping windows of 10 kb across each entire chromosome. Tajima's *D* is a measure of deviation from neutral evolution, computed as the difference between the mean number of pairwise differences and the number of segregating sites. Values less than −2 or greater than two are generally considered as strong indication that a gene (or a genomic region) is not evolving neutrally. Genes with a Tajima's *D* value below −2 have an excess of rare alleles, indicating positive selection or a selective sweep, whereas genes with a Tajima's *D* value greater than two have an excess of common alleles suggestive of balancing selection.

Genetic diversity within and between populations, namely, nucleotide diversity (π) and absolute divergence (d_{xy}), were calculated in genomic windows of 50 kb, sliding by 25 kb (https://github.com/simonhmartin/genomics_general). Briefly, π is the average number of nucleotide differences between genotypes drawn from the same population, whereas d_{XY} is the average number of nucleotide differences between genotypes drawn from two different populations.

The fixation index (F_{st}) for each population was computed in windows of 1 kb (https://github.com/simonhmartin/genomics_general).

Decay in LD was estimated with PopLDdecay 3.42 (Zhang et al. 2019), measuring r^2 between SNPs until 300 kb. The values were computed comparing the mean values of 100 pseudoreplicates, each one composed by 10 isolates extracted randomly.

Selective pressure analyses

The phylogenetic tree inferred from genomic SNPs was labeled according to the population structure using the dedicated tool of Hyphy v.2.5.50 (Murrell et al. 2015). Then, we determined whether a gene was subjected to positive selection using Hyphy with the BUSTED algorithm on the respective gene sequences from the reconstructed chromosomes (see above). Genes with a *P*-value < 0.05 were considered statistically significant.

Comparison of putative virulence genes

A set of 55 putative virulence genes (Dumaine et al. 2021) was retrieved. These genes include members of small gene families characterized by possessing specific protein domains (MEDLE, WYLE, GGC, FLGN, SKSR, and mucins) and by having N-terminal signal peptides. The corresponding protein sequences were identified in the assembly of each isolate using BLAST. The results were manually curated, and multiple protein alignments were generated. The presence and distribution pattern of amino acid substitutions were investigated manually.

Data access

The raw sequence data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA885600. All relevant codes used to generate the images are available as Supplemental Code S1 and at GitHub (https://github.com/MIDIfactory/cryptosporidium_parvum_evolution).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Dr. Christian Seyboldt (Friedrich Loeffler Institute, Germany) and Dr. Ernst Grossmann (Aulendorf State Veterinary diagnostic centre, Germany) for generous provision of original samples. We also thank Elke Radam (Robert Koch Institute, Germany) for excellent technical assistance and Dr. Giulia Colombo (University of Pavia, Italy) for helpful discussion. This work was supported by the European Union's Horizon 2020 Research and Innovation program under grant agreement no. 773830: One Health European Joint Programme, PARADISE project (https://onehealthejp.eu/jrp-paradise/). T.E. and R.R.-F. were funded by the Development Fund Agricultural and Forestry MAKERA, Ministry of Agriculture and Forestry of Finland (grant no. 435/03.01.02/2018).

Author contributions: S.M.C. conceived the study. G.B., T.N., M.C., and G.B.B. performed the bioinformatics analyses. A.R.S. and S.M.C. performed the bench work. S.M.C., C.B., G.B., T.N., M.C., and D.S. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

References

Alves M, Xiao L, Sulaiman I, Lal AA, Matos O, Antunes F. 2003. Subgenotype analysis of *Cryptosporidium* isolates from humans, cattle, and zoo ruminants in Portugal. *J Clin Microbiol* **41:** 2744–2747. doi:10.1128/JCM.41.6.2744-2747.2003

Baptista RP, Li Y, Sateriale A, Sanders MJ, Brooks KL, Tracey A, Ansell BRE, Jex AR, Cooper GW, Smith ED, et al. 2022. Long-read assembly and comparative evidence-based reanalysis of *Cryptosporidium* genome sequences reveal expanded transporter repertoire and duplication of entire chromosome ends including subtelomeric regions. *Genome Res* **32:** 203–213. doi:10.1101/gr.275325.121

Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, et al. 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial

communities with bioBakery 3. *eLife* **10:** e65088. doi:10.7554/eLife.65088
Beja-Pereira A, Caramelli D, Lalueza-Fox C, Vernesi C, Ferrand N, Casoli A, Goyache F, Royo LJ, Conti S, Lari M, et al. 2006. The origin of European cattle: evidence from modern and ancient DNA. *Proc Natl Acad Sci* **103:** 8113–8118. doi:10.1073/pnas.0509210103
Bhalchandra S, Cardenas D, Ward HD. 2018. Recent breakthroughs and ongoing limitations in *Cryptosporidium* research. *F1000Res* **7:** F1000 Faculty Rev-1380. doi:10.12688/f1000research.15333.1
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120. doi:10.1093/bioinformatics/btu170
Bowling GA. 1942. The introduction of cattle into colonial North America. *J Dairy Sci* **25:** 129–154. doi:10.3168/jds.S0022-0302(42)95275-5
Braima K, Zahedi A, Oskam C, Reid S, Pingault N, Xiao L, Ryan U. 2019. Retrospective analysis of *Cryptosporidium* species in Western Australian human populations (2015–2018), and emergence of the *C. hominis* IfA12G1R5 subtype. *Infect Genet Evol* **73:** 306–313. doi:10.1016/j.meegid.2019.05.018
Cacciò SM, Chalmers RM. 2016. Human cryptosporidiosis in Europe. *Clin Microbiol Infect* **22:** 471–480. doi:10.1016/j.cmi.2016.04.021
Chavez MA, White C Jr. 2018. Novel treatment strategies and drugs in development for cryptosporidiosis. *Expert Rev Anti Infect Ther* **16:** 655–661. doi:10.1080/14787210.2018.1500457
Chessa B, Pereira F, Arnaud F, Amorim A, Goyache F, Mainland I, Kao RR, Pemberton JM, Beraldi D, Stear MJ, et al. 2009. Revealing the history of sheep domestication using retrovirus integrations. *Science* **324:** 532–536. doi:10.1126/science.1170587
Corsi GI, Tichkule S, Sannella AR, Vatta P, Asnicar F, Segata N, Jex AR, van Oosterhout C, Cacciò SM. 2023. Recent genetic exchanges and admixture shape the genome and population structure of the zoonotic pathogen *Cryptosporidium parvum*. *Mol Ecol* **32:** 2633–2645. doi:10.1111/mec.16556
Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems*: 1695. https://igraph.org.
Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27:** 2156–2158. doi:10.1093/bioinformatics/btr330
Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10:** giab008. doi:10.1093/gigascience/giab008
Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol Biol Evol* **37:** 291–294. doi:10.1093/molbev/msz189
Delsol N, Stucky BJ, Oswald JA, Cobb CR, Emery KF, Guralnick R. 2023. Ancient DNA confirms diverse origins of early post-Columbian cattle in the Americas. *Sci Rep* **13:** 12444. doi:10.1038/s41598-023-39518-3
DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43:** 491–498. doi:10.1038/ng.806
Dumaine JE, Sateriale D, Gibson AR, Reddy AG, Gullicksrud JA, Hunter EN, Clark JT, Striepen B. 2021. The enteric pathogen *Cryptosporidium parvum* exports proteins into the cytosol of the infected host cell. *eLife* **10:** e70451. doi:10.7554/eLife.70451
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792–1797. doi:10.1093/nar/gkh340
Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20:** 238. doi:10.1186/s13059-019-1832-y
Feng Y, Li N, Roellig DM, Kelley A, Liu G, Amer S, Tang K, Zhang L, Xiao L. 2017. Comparative genomic analysis of the IId subtype family of *Cryptosporidium parvum*. *Int J Parasitol* **47:** 281–290. doi:10.1016/j.ijpara.2016.12.002
Feng Y, Ryan UM, Xiao L. 2018. Genetic diversity and population structure of *Cryptosporidium*. *Trends Parasitol* **34:** 997–1011. doi:10.1016/j.pt.2018.07.009
Ficek RE. 2019. Cattle, capital, colonization. tracking creatures of the Anthropocene in and out of human projects. *Current Anthropol* **60:** S260–S271. doi:10.1086/702788
Garcia-R JC, Hayman DTS. 2016. Origin of a major infectious disease in vertebrates: the timing of *Cryptosporidium* evolution and its hosts. *Parasitology* **143:** 1683–1690. doi:10.1017/S0031182016001323
Gruber-Vodicka HR, Seah BKB, Pruesse E. 2020. phyloFlash: rapid small-subunit rRNA profiling and targeted assembly from metagenomes. *mSystems* **5:** e00920-20. doi:10.1128/mSystems.00920-20
Guo Y, Ryan U, Feng Y, Xiao L. 2022. Association of common zoonotic pathogens with concentrated animal feeding operations. *Front Microbiol* **12:** 810142. doi:10.3389/fmicb.2021.810142
Hadfield SJ, Pachebat JA, Swain MT, Robinson G, Cameron S, Alexander J, Hegarty MJ, Elwin K, Chalmers RM. 2015. Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples. *BMC Genomics* **16:** 650. doi:10.1186/s12864-015-1805-9
Hemstrom W, Jones M. 2023. snpR: user friendly population genomics for SNP data sets with categorical metadata. *Mol Ecol Res* **23:** 962–973. doi:10.1111/1755-0998.13721
Hijjawi N, Zahedi A, Al-Falah M, Ryan U. 2022. A review of the molecular epidemiology of *Cryptosporidium* spp. and *Giardia duodenalis* in the Middle East and North Africa (MENA) region. *Infect Genet Evol* **98:** 105212. doi:10.1016/j.meegid.2022.105212
Huang W, Guo Y, Lysen C, Wang Y, Tang K, Seabolt MH, Yang F, Cebelinski E, Gonzalez-Moreno O, Hou T, et al. 2023. Multiple introductions and recombination events underlie the emergence of a hyper-transmissible *Cryptosporidium hominis* subtype in the USA. *Cell Host Microbe* **31:** 112–123.e4. doi:10.1016/j.chom.2022.11.013
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23:** 254–267. doi:10.1093/molbev/msj030
Innes EA, Chalmers RM, Wells B, Pawlowic MC. 2020. A one health approach to tackle cryptosporidiosis. *Trends Parasitol* **36:** 290–303. doi:10.1016/j.pt.2019.12.016
Jann HW, Cabral-Castro MG, Barreto Costa JV, de Barros Alencar ACM, Peralta JM, Saramago Peralta RE. 2022. Prevalence of human cryptosporidiosis in the Americas: systematic review and meta-analysis. *Rev Inst Med Trop Sao Paulo* **64:** e70. doi:10.1590/S1678-9946202264070
Khalil IA, Troeger C, Rao PC, Blacker BF, Brown A, Brewer TG, Colombara DV, De Hostos EL, Engmann C, Guerrant RL, et al. 2018. Morbidity, mortality, and long-term consequences associated with diarrhoea from *Cryptosporidium* infection in children younger than 5 years: a meta-analyses study. *Lancet Glob Health* **6:** e758–e768. doi:10.1016/S2214-109X(18)30283-3
Khan SM, Witola WH. 2023. Past, current, and potential treatments for cryptosporidiosis in humans and farm animals: a comprehensive review. *Front Cell Infect Microbiol* **13:** 1115522. doi:10.3389/fcimb.2023.1115522
Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, Wu Y, Sow SO, Sur D, Breiman RF, et al. 2013. 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. doi:10.1016/S0140-6736(13)60844-2
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359. doi:10.1038/nmeth.1923
Long S, Anthony B, Drewry LL, Sibley LD. 2017. A conserved Ankyrin repeat-containing protein regulates conoid stability, motility and cell invasion in *Toxoplasma gondii*. *Nat Commun* **8:** 2236. doi:10.1038/s41467-017-02341-2
Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, O'Brien J, Djimde A, Doumbo O, Zongo I, et al. 2012. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* **487:** 375–379. doi:10.1038/nature11174
Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol* **1:** vev003. doi:10.1093/ve/vev003
Mathur V, Wakeman KC, Keeling PJ. 2021. Parallel functional reduction in the mitochondria of apicomplexan parasites. *Curr Biol* **31:** 2920–2928. doi:10.1016/j.cub.2021.04.028
McKerr C, O'Brien SJ, Chalmers RM, Vivancos R, Christley RM. 2018. Exposures associated with infection with *Cryptosporidium* in industrialised countries: a systematic review protocol. *Syst Rev* **7:** 70. doi:10.1186/s13643-018-0731-8
McTavish EJ, Decker JE, Schnabel RD, Taylor JF, Hillis DH. 2013. New World cattle show ancestry from multiple independent domestication events. *Proc Natl Acad Sci* **110:** E1398–E1406. doi:10.1073/pnas.1303367110
Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, et al. 2015. Gene-wide identification of episodic selection. *Mol Biol Evol* **32:** 1365–1371. doi:10.1093/molbev/msv035
Nader JL, Mathers TC, Ward BJ, Pachebat JA, Swain MT, Robinson G, Chalmers RM, Hunter PR, van Oosterhout C, Tyler KM. 2019. Evolutionary genomics of anthroponosis in *Cryptosporidium*. *Nat Microbiol* **4:** 826–836. doi:10.1038/s41564-019-0377-x
Nash JHE, Robertson J, Elwin K, Chalmers RA, Kropinski AM, Guy RA. 2018. Draft genome assembly of a potentially zoonotic *Cryptosporidium parvum* isolate, UKP1. *Microbiol Resour Announc* **7:** e01291-18. doi:10.1128/MRA.01291-18

Peake L, Inns T, Jarvis C, King G, Rabie H, Henderson J, Wensley A, Jarratt R, Roberts C, Williams C, et al. 2023. Preliminary investigation of a significant national *Cryptosporidium* exceedance in the United Kingdom, August 2023 and ongoing. *Euro Surveill* **28:** 2300538. doi:10.2807/1560-7917.ES.2023.28.43.2300538

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959. doi:10.1093/genetics/155.2.945

Puiu D, Enomoto S, Buck GA, Abrahamsen MS, Kissinger JC. 2004. CryptoDB: the *Cryptosporidium* genome resource. *Nucleic Acids Res* **32:** D329–D331. doi:10.1093/nar/gkh050

Rahman SU, Mi R, Zhou S, Gong H, Ullah M, Huang Y, Han X, Chen Z. 2022. Advances in therapeutic and vaccine targets for *Cryptosporidium*: challenges and possible mitigation strategies. *Acta Trop* **226:** 106273. doi:10.1016/j.actatropica.2021.106273

R Core Team. 2021. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Ryan U, Xiao L, Read C, Zhou L, Lal AA, Pavlasek I. 2003. Identification of novel *Cryptosporidium* genotypes from the Czech Republic. *Appl Environ Microbiol* **69:** 4302–4307. doi:10.1128/AEM.69.7.4302-4307.2003

Ryan UM, Feng Y, Fayer R, Xiao L. 2021a. Taxonomy and molecular epidemiology of *Cryptosporidium* and *Giardia* – a 50 year perspective (1971–2021). *Int J Parasitol* **51:** 1099–1119. doi:10.1016/j.ijpara.2021.08.007

Ryan U, Zahed A, Feng Y, Xiao L. 2021b. An update on zoonotic *Cryptosporidium* species and genotypes in humans. *Animals (Basel)* **11:** 3307. doi:10.3390/ani11113307

Schaffner SF, Taylor AR, Wong W, Wirth DF, Neafsey DE. 2018. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *Malaria J* **17:** 196. doi:10.1186/s12936-018-2349-7

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313. doi:10.1093/bioinformatics/btu033

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34:** W435–W439. doi:10.1093/nar/gkl200

Straschil U, Talman AM, Ferguson DJP, Bunting KA, Xu Z, Bailes E, Sinden RE, Holder AA, Smith EF, Coates JC, et al. 2010. The Armadillo repeat protein pf16 is essential for flagellar structure and function in *Plasmodium* male gametes. *PLoS One* **5:** e12901. doi:10.1371/journal.pone.0012901

Troell K, Hallström B, Divne A, Alsmark C, Arrighi R, Huss M, Beser J, Bertilsson S. 2016. *Cryptosporidium* as a testbed for single cell genome characterization of unicellular eukaryotes. *BMC Genomics* **17:** 471. doi:10.1186/s12864-016-2815-y

Tůmová L, Ježková J, Prediger J, Holubová N, Sak B, Konečný R, Květoňová D, Hlásková L, Rost M, McEvoy J, et al. 2023. *Cryptosporidium mortiferum* n. sp. (Apicomplexa: Cryptosporidiidae), the species causing lethal cryptosporidiosis in Eurasian red squirrels (*Sciurus nulgaris*). *Parasites Vect* **16:** 235. doi:10.1186/s13071-023-05844-8

Van der Auwera GA, Connor O, D B. 2020. *Genomics in the cloud: using docker, GATK, and WDL in terra*. O'Reilly Media, Sebastopol, CA.

Wang R, Zhang L, Axén C, Bjorkman C, Jian F, Amer S, Liu A, Feng Y, Li G, Lv C, et al. 2014. *Cryptosporidium parvum* IId family: clonal population and dispersal from Western Asia to other geographical regions. *Sci Rep* **4:** 4208. doi:10.1038/srep04208

Wang T, Guo Y, Roellig DM, Li N, Santín M, Lombard J, Kváč M, Naguib D, Zhang Z, Feng Y, et al. 2022. Sympatric recombination in zoonotic *Cryptosporidium* leads to emergence of populations with modified host preference. *Mol Biol Evol* **39:** msac150. doi:10.1093/molbev/msac150

Wick R, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* **13:** e1005595. doi:10.1371/journal.pcbi.1005595

Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. Springer, Cham, Switzerland.

Xu Z, Guo Y, Roellig DM, Feng Y, Xiao L. 2019. Comparative analysis reveals conservation in genome organization among intestinal *Cryptosporidium* species and sequence divergence in potential secreted pathogenesis determinants among major human-infecting species. *BMC Genomics* **20:** 406. doi:10.1186/s12864-019-5788-9

Zahedi A, Ryan U. 2020. *Cryptosporidium*: an update with an emphasis on foodborne and waterborne transmission. *Res Vet Sci* **132:** 500–512. doi:10.1016/j.rvsc.2020.08.002

Zhang C, Dong S, Xu J, He W, Yang T. 2019. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35:** 1786–1788. doi:10.1093/bioinformatics/bty875

Received December 11, 2023; accepted in revised form June 5, 2024.

Comparative genomics of *Cryptosporidium parvum* reveals the emergence of an outbreak-associated population in Europe and its spread to the United States

Greta Bellinzona, Tiago Nardi, Michele Castelli, et al.

Genome Res. 2024 34: 877-887 originally published online July 8, 2024
Access the most recent version at doi:10.1101/gr.278830.123