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

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

1 Department of Biology and Biotechnology, University of Pavia, Italy; 2 Department of Comparative Biomedical Sciences, School of Veterinary Medicine, University of Surrey, Guildford, UK; 3 Viral Genetics and Biosecurity Unit (GVB), French Agency for Food, Environmental and Occupational Health Safety (ANSES), Ploufragan, France; 4 Department of Microbiology, Public Health Agency of Sweden, Solna, Sweden; 5 Cryptosporidium Reference Unit, Swansea, UK; 6 Norwegian Veterinary Institute, Ås, Norway; 7 Istituto Zooprofilattico Sperimentale dell'Umbria e delle Marche, Perugia, Italy; 8 Animal Health Diagnostic Unit, Finnish Food Authority, Kuopio, Finland; 9 National Institute for Agricultural and Veterinary Research, Lisbon, Portugal; 10 Ecole Nationale Vétérinaire d'Alfort, Laboratoire de Santé Animale, Maisons-Alfort, France; 11 Department of Infectious Diseases, Unit for Mycotic and Parasitic Agents and Mycobacteria, Robert Koch Institute, Berlin, Germany; 12 National Veterinary Institute, Uppsala, Sweden; 13 National Institute for Public Education, Budapest, Hungary; 14 Finnish Institute for Health and Welfare, Helsinki, Finland; 15 Department of Infectious Diseases, Istituto Superiore di Sanità, Rome, Italy; 16 Department of Parasitology and Invasive Diseases, National Veterinary Research Institute, Pulawy, Poland; 17 Statens Serum Institut, Copenhagen, Denmark; 18 Veterinary Research Institute, Department of Food and Feed Safety, Brno, Czech Republic; 19 Department of Biosciences, University of Milan, Milan, Italy; 20 IRCCS Fondazione Policlinico San Matteo, Pavia, Italy

* corresponding authors: davide.sassera@unipv.it; simone.caccio@iss.it

Abstract

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 USA (n=72) in the largest comparative genomics study to date. We applied rigorous filters to exclude mixed infections and analysed a dataset 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, while the majority of European (IIa, including all IIaA15G2R1 isolates) and all isolates from the USA (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 UK, reach the USA where it also expanded. In addition, genetic exchanges between different populations led to the formation of mosaic genomes. The reason(s) for the successful spread of population 3 remained elusive, although genes under selective pressure uniquely in this population were identified.

Introduction

The genus *Cryptosporidium* (phylum Apicomplexa) currently comprises 46 species and more than 120 genotypes of uncertain taxonomic status (Innes et al. 2020; U. M. Ryan et al. 2021; 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 diarrhoea (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 (U. Ryan et al. 2021). Indeed, calibrated phylogenies indicate that much of *Cryptosporidium* 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 *C. parvum*, *C. felis, C. canis, C. cuniculus, C. ubiquitum, C. meleagridis* and others (Rachel M. Chalmers et al. 2018; 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, Cardenas, and Ward 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 due to *C. hominis*, which is anthroponotic, or *C. parvum*, which is zoonotic (Feng, Ryan, and Xiao 2018). Animal reservoirs, in particular young ruminants, have an essential role in the spillover and spillback of *C. parvum* to humans (Guo et al. 2021).

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, Ryan, and Xiao 2018). In Europe, many IIa subtypes have been identified in humans, and many also circulate among animals. However, a few subtypes appear to predominate, particularly subtype IIaA15G2R1, which is also the most common subtype globally (R. M. Chalmers and Cacciò 2016). The reasons for this high prevalence are unknown.

Recent studies based on whole genome sequence (WGS) comparisons have started to explore the evolutionary genetics of *C. parvum* (Corsi et al. 2023; T. Wang et al. 2022; Feng et al. 2017). 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 these populations. Another work analysed 101 WGSs and hypothesised 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 (T. Wang et al. 2022).

In this study, we generated WGSs for 123 human- or ruminant-derived *C. parvum* isolates collected across Europe, and retrieved publicly available WGS data of 71 isolates from Europe, Egypt, China and the USA (Corsi et al. 2023; T. Wang et al. 2022; Feng et al. 2017; Hadfield et al. 2015; Troell et al. 2016). 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 USA.

Material and Methods

1. Parasite isolates

Table S1 lists the information available for the 195 *C. parvum* isolates from humans and ruminants included in this study. The dataset 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; Corsi et al. 2023; T. Wang et al. 2022), and the recently assembled IOWA-ATCC genome (Baptista et al. 2022), which was used as a reference genome.

2. Oocyst purification, DNA processing and sequencing

An aliquot of the 123 faecal isolates was used to extract genomic DNA and to identify the species and the *gp60* subtype, using previously published protocols (U. Ryan et al. 2003; Alves et al. 2003). The procedures for DNA purification and extraction are detailed in Corsi et al. (Corsi et al. 2023). In short, oocyst were purified from faecal 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 Midi-Kit (Qiagen), according to the manufacturer's instructions.

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

3. Data filtering and SNP calling

Raw reads of the 194 isolates were quality-checked and then pre-processed to remove low-quality bases and adapter sequences using Trimmomatic v.0.36 (Bolger, Lohse, and Usadel 2014), with default parameters. A series of sequential steps were then applied to select isolates and SNPs according to multiple criteria (Supplementary Figure 1, Supplementary 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, Seah, and Pruesse 2020). Only isolates showing the presence of *Cryptosporidium* spp. were retained for further analyses.

The *C. parvum* IOWA-ATCC (Baptista et al. 2022) was used as a reference genome to map the filtered reads of each sample with bowtie2 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 option -ERC GVCF. SNPs were removed if quality depth <2.0, Fisher strand >60.0, mapping quality <30.0, mapping quality rank-sum test <−12.5, read position rank-sum test <−8.0, and strand odds ratio >3.0.

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

To maximise the quality, SNPs were further filtered using bcftools based on the following criteria: biallelic SNPs, quality score >30, allele depth >20, minor allele frequency >0.005, and missing ratio <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, FWS is expected to approach unity. Isolates with 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 =1 or >1 are presented in Supplementary Figure 2.

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 dataset (Supplementary Table 1).

SNP differences were calculated using snp-dists v0.8.2 (https://github.com/tseemann/snp-dists) and visualised 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 visualised using the R package ggplot2 (Wickham 2016).

To compare the SNP density between each chromosome we used the pairwise comparisons for proportions test implemented in R, and the probability (p) values were adjusted using the Bonferroni correction.

4. 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 dataset, and used *Cryptosporidium 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 all the remaining 140 isolates. A set of 195 genes, which have been used previously for phylogenomic analyses of Apicomplexa (Mathur, Wakeman, and Keeling 2021), was searched using BLASTp on the orthogroups identified by OrthoFinder v2.5.4 (Emms and Kelly 2019) in our dataset. Of these, orthologs of 179 genes were identified. Each ortholog was aligned with Muscle 5.1 (Edgar 2004) and concatenated. A maximum likelihood (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 pseudo-replicates.

Next, a concatenated set of *C. parvum* 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 pseudo-replicates. 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, Stephens, and Donnelly 2000), with the number of populations tested (K) ranging from 1 to 12. Phylogenetic networks were generated by using the Neighbor-Net algorithm implemented in SplitsTree v.5 (Huson and Bryant 2006).

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

5. 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 analysed by the Recombination Detection Program software, version 5 (RDP5) (Murrell 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 cut-off of 10e-5 were considered significant.

6. Population Genetic Analyses

Tajima's D values were calculated using snpR (Hemstrom and Jones 2023) in non-overlapping windows of 10 Kb across each entire chromosome. Pairwise divergence (Dxy) and intra-population nucleotide diversity (π) were calculated in genomic windows of 50 kbp sliding by 25 kbp (https://github.com/simonhmartin/genomics_general). The Fixation index (Fst) for each population was computed in windows of 1 kb (https://github.com/simonhmartin/genomics_general).

Decay in Linkage Disequilibrium (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.

7. Selective pressure analyses

The phylogenetic tree inferred from genomic SNPs was labelled 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.

8. 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 characterised 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.

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). In order to get a robust foundation for reliable inferences, we performed a careful selection based on multiple criteria, including the level of contamination from non-target organisms, mean read depth, multiplicity of infection, and genome assembly quality (more details are provided in the Materials and Methods section and in Supplementary Table S1). This stringent selection process yielded a final dataset of 141 isolates (including 88 newly sequenced isolates and 52 publicly available isolates), and the reference IOWA-ATCC, which were used for downstream analyses.

Importantly, the final dataset 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 dataset composition can be found in the Supplementary Table S1.

Genetic variability among isolates

By comparing the 140 isolates to the IOWA-ATCC reference, a total of 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 SNPs distribution was non-random at the level of individual chromosomes, with statistically significant higher SNP density observed at chromosomes 1 and 6 (p-value < e-05) (Supplementary Table S2), and with an enrichment in subtelomeric regions compared with internal regions of the chromosomes (Supplementary Figure 3).

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 (Supplementary Figure 4). We observed that all isolates from Europe and the USA formed a highly supported, monophyletic clade (hereafter, the "Western" lineage), whereas isolates from China and Egypt appeared to have diverged earlier.

Next, to get a finer description of the relationships among the 141 *C. parvum* isolates, we inferred a ML tree based on the concatenated set of 28,047 biallelic SNPs (Figure 1), using the root determined in the tree based on orthologous genes. The large-scale topology was consistent with the orthologous genes-based phylogeny. Corroborating indications from previous studies (Corsi et al. 2023; T. Wang et al. 2022), 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, 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). Interestingly, all isolates from the USA formed a fully supported monophyletic clade that was nested within European isolates and showed an intriguing sister group relationship with a clade of isolates from the UK.

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

While the three populations were genetically very distinct (Figure 1), admixed isolates were also evident, most notably the IIa isolates from Egypt, and in the European isolates FIN3, GER13 and DK8. A phylogenetic network showed several connections between isolates from different populations, suggestive of recombination events (Supplementary Figure 5).

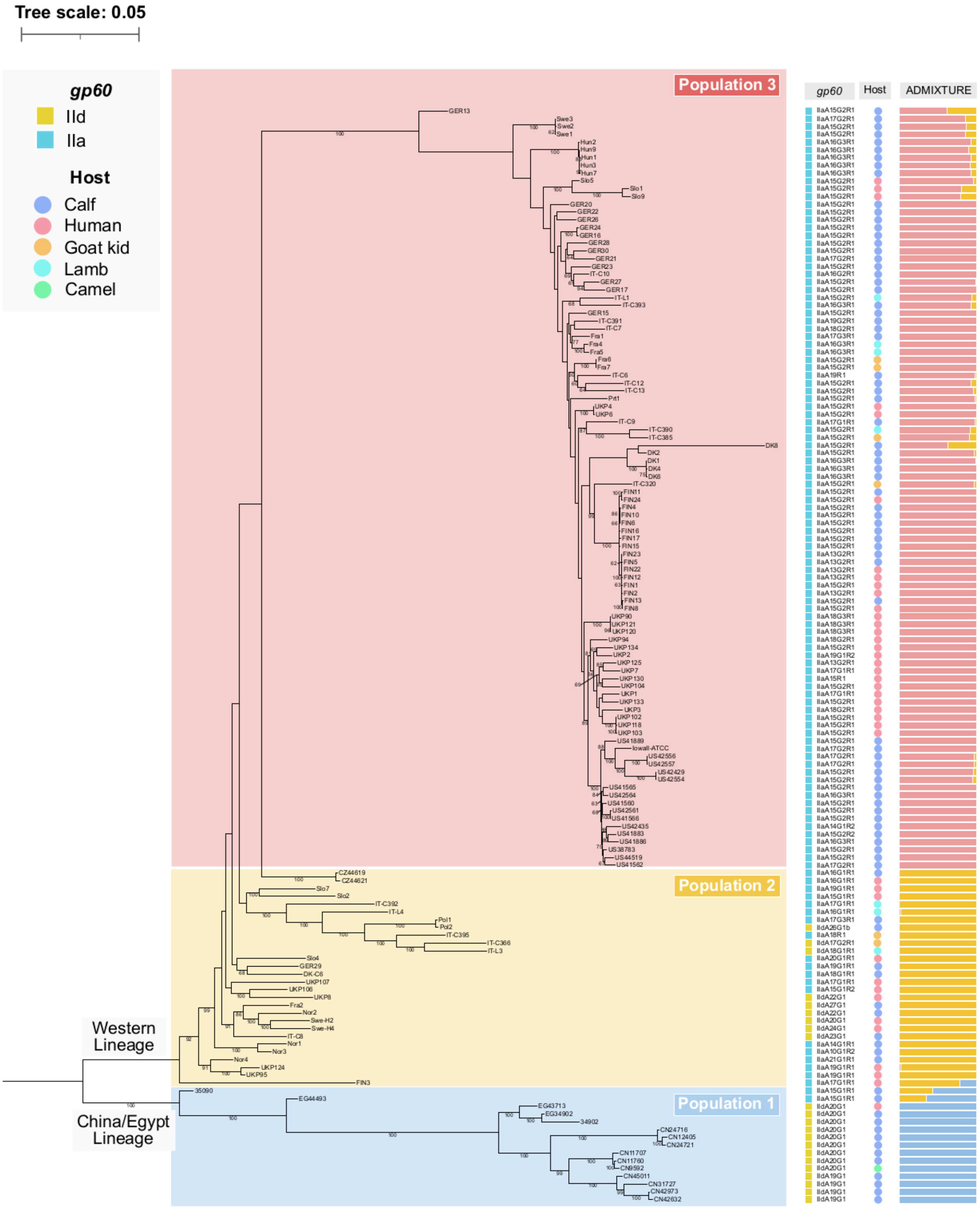

Figure 1. Maximum Likelihood tree based on a set of 28,047 biallelic SNPs. Only bootstrap values >60 are shown. Information about host, *gp60* subtype and results of ADMIXTURE are mapped on the phylogeny.

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), ADMIXTURE (Zhou, Alexander, and Lange 2011), and SplitsTree (Huson and Bryant 2006). RDP5 (Martin et al. 2015) analyses were performed to provide statistical evidence for putative recombination events.

At chromosome 1, phylogenetic analysis showed the IIa isolates from Egypt (35909 and EG4493) cluster with population 2 and not with population 1, in contrast with the topology based on all genomic SNPs. 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). Indeed, in the region spanning position 755,934 to 768,672 (about 15 kb), 260 SNPs are found in population 1 (including isolates 35909 and EG4493), while populations 2 and 3 have very limited genetic variability. Immediately after this block, the IIa isolates from Egypt are essentially identical to those from population 2 until position 823,729 (about 55 kb), while the IId isolates differ from the reference genome by 560 SNPs in this region. 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.

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. Indeed, in a 50 kb region spanning position 824,800 to 874,170, the isolate FIN3 shared 443 SNPs with the IId isolates from population 1. 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.

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. 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. 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 differ from all other isolates of populations 2 and 3, which are 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, while all other population 3 isolates are essentially invariant, being identical to the reference genome. 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.

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, while the remaining isolates from population 3 are largely invariant. Five genes are located in this region, and encode for 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 dataset of 140 genomes, only 1,243 (4.4%) were shared by the three populations, while the majority was specific to a single population (Supplementary Figure 6).

We calculated the pairwise SNP distances among the 141 isolates and observed the smallest distances in population 3 (range 3 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 5,867 and 5,241 SNPs, respectively, while population 2 and population 3 displayed a lower average SNP distance (3,847 SNPs).

Furthermore, we observed 13 clusters of highly similar genomes (defined as having < 50 SNPs) (Figure 2) all belonging to population 3, encompassing from 2 to 8 isolates. Notably, all 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, while isolates UKP90, UKP120 and UKP121 were from a distinct outbreak that occurred in April 2016. Another cluster of highly similar genomes (i.e., differing for < 20 SNPs) was formed by six Hungarian calf isolates (Hun1, Hun2, Hun3, Hun7 and Hun9), collected at a single farm from the 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, a very high genomic similarity was observed (pairwise SNP distance < 20 SNPs), and the respective isolates formed monophyletic, highly supported clusters in the phylogenetic analysis (Figure 1). Although the isolates 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 cut-off to confirm suspected outbreaks at the genomic level.

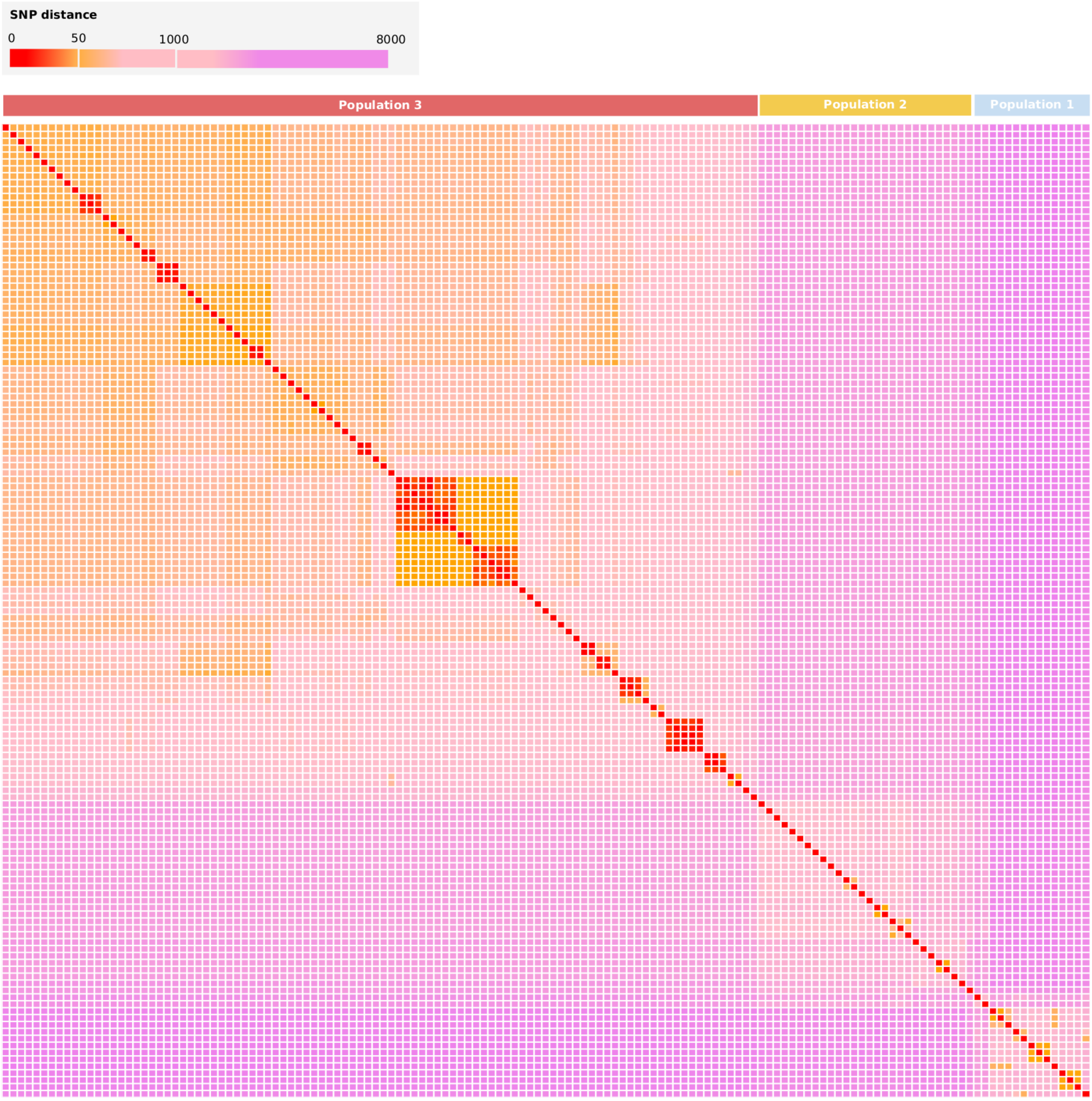

Figure 2. Heatmap illustrating the SNP distance among the 141 isolates analysed. The order of the isolates reflects the position they occupy in the SNP-based phylogenetic tree. The colour code is shown in the legend on the top.

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., where the fraction of shared IBD is greater than 90% or 80%). As shown in Supplementary Figure 7, 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 populations. Notably, networks were observed within population 3 (Hungary, Finland, USA/UK, 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., population 2 and 3), we calculated nucleotide diversity (π), linkage disequilibrium (LD) decay, and Tajima's D.

Assessing nucleotide diversity (π) within each population unveiled notable distinctions. Population 3 exhibited lower diversity compared to population 2 both at the entire genome level (respective means $\pi = 0.039$ and $\pi = 0.073$) and also when analysing single chromosomes (Figure 3). This reduced genetic variation is consistent with a recent origin of population 3, and can be explained by various factors, such as genetic drift (e.g., population bottlenecks) or selective sweeps.

We then examined LD decay and found that population 3 had a slower decay than population 2 (Figure 3), reinforcing the concept of a more recent origin of population 3.

Finally, we observed that the distribution of Tajima's D values in population 3 is skewed towards negative values (Figure 3), indicating an excess of rare polymorphisms. This skewness was less pronounced in population 2 (Figure 3).

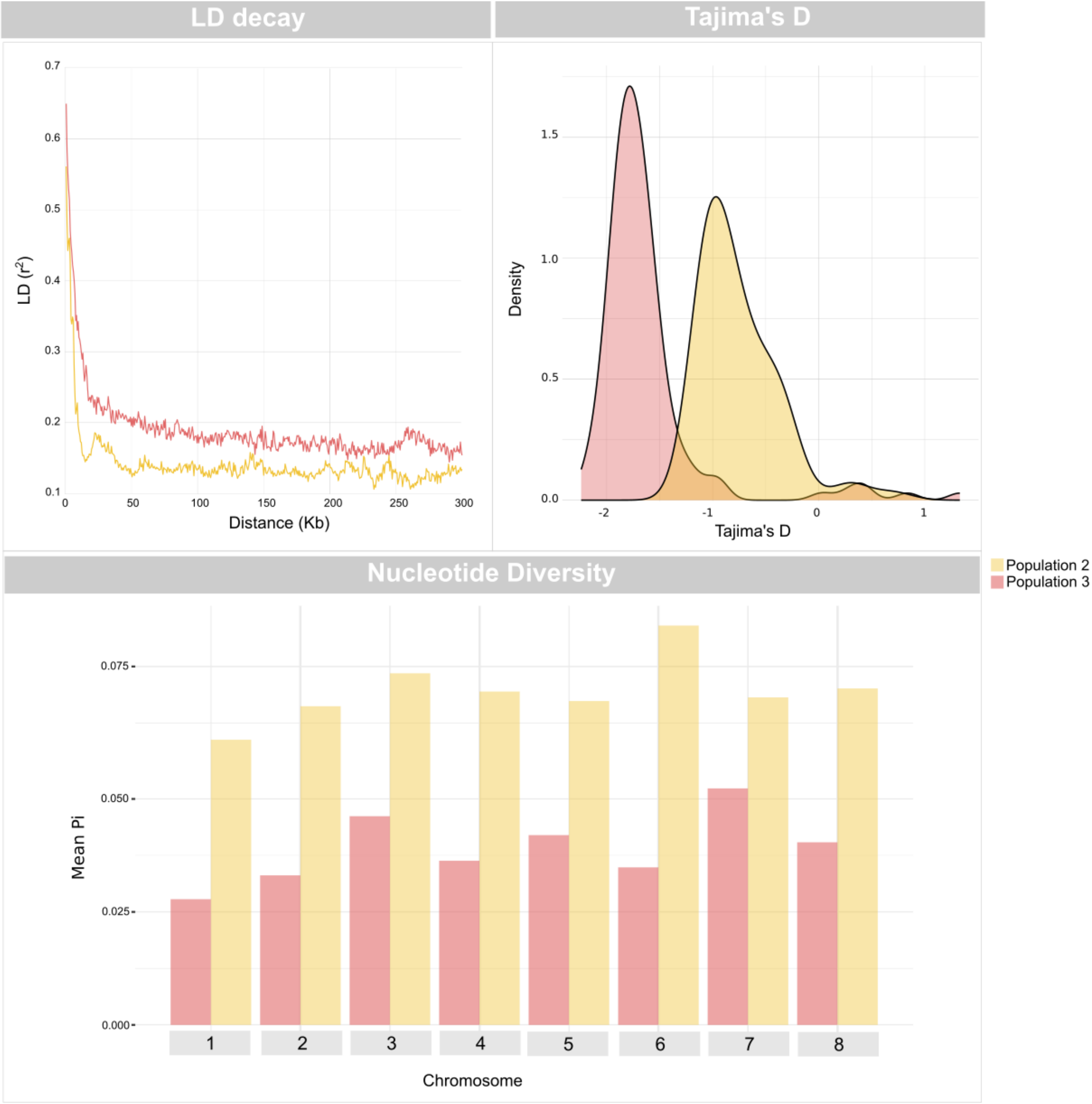

Figure 3. LD decay, distribution of Tajima's D values and nucleotide diversity (π) in population 2 and population 3. Nucleotide diversity is shown with the mean Pi for each chromosome. Chromosome-specific results are provided in Supplementary Figure 8.

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, neither considering amino acid substitutions.

To investigate differences between population 3 and population 2 at the genome-wide level, we calculated the fixation index (Fst) in 1 kb windows (Figure 4). This allowed 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 Fst values (i.e., applying a cut-off of 0.91), we identified 79 regions (Figure 4) and highlighted candidate genes under selection according to their function in Figure 4 (Supplementary Table 3).

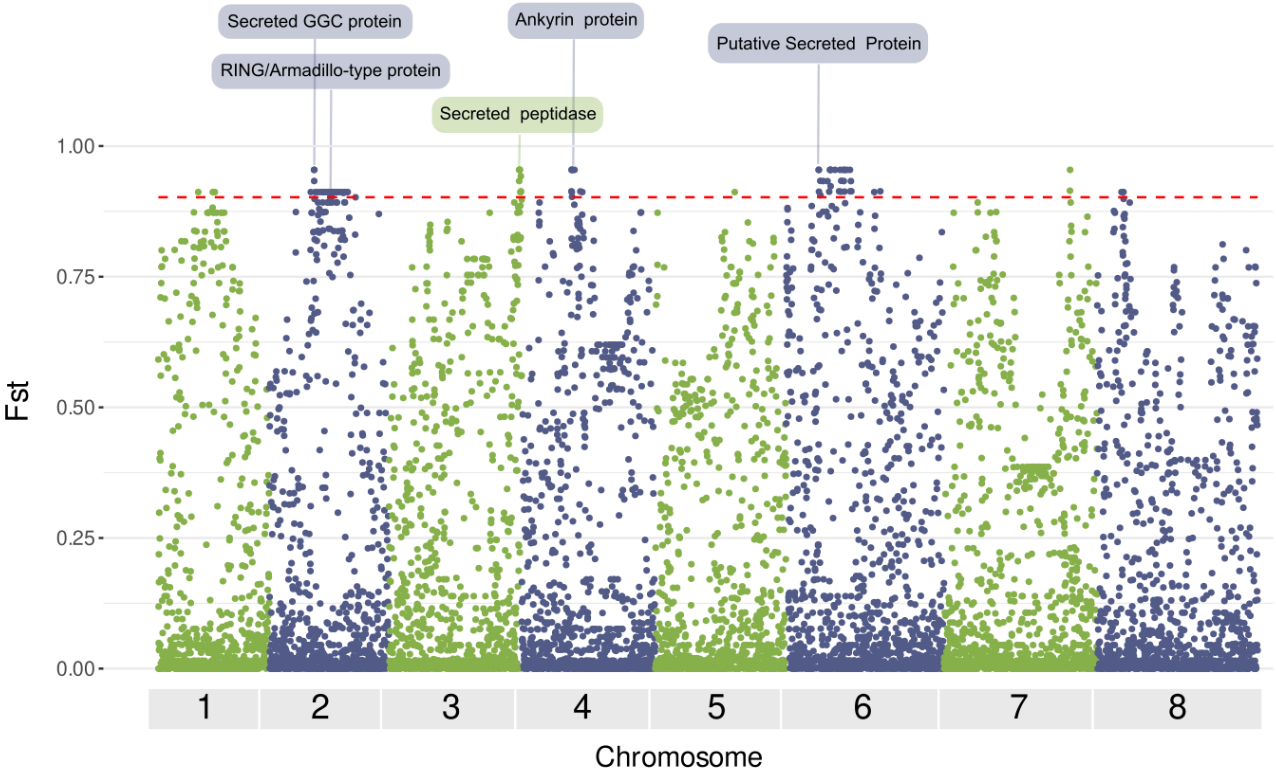

Figure 4. Manhattan plot of genome-wide Wright's Fst values, calculated in genomic regions of 1 kb, comparing population 2 and population 3. Fst values are shown on the y axis, and genomic positions on the x axis. The dotted red line represents the cut-off value of the top 1%, equal to 0.91. Genes overlapping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

Then, 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," while a few (8/176) were "putative secreted proteins" (Supplementary Table S3).

Notably, 16 proteins were identified in both the analyses, i.e., Fst statistics on genomic regions and selective pressure in single genes (Supplementary Table S3).

Discussion

C. parvum is the most prevalent zoonotic pathogen within the genus *Cryptosporidium*, and a global cause of diarrheal disease in humans and ruminants (Kotloff et al. 2013). This species is widespread in industrialised countries, including Europe (Cacciò and Chalmers 2016), but also in the Middle East (Hijjawi et al., 2022). Despite the recognised 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; Khan and Witola 2023; Rahman et al. 2022). Recent WGS studies have started to provide insights into the genetics of *C. parvum*, proposing a role of recombination events in the evolution of this species and have allowed to identify a number of genes under positive selection, potentially involved in host-parasite interactions (Corsi et al. 2023; T. Wang et al. 2022).

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 USA (n=71) were included (Hadfield et al. 2015; Troell et al. 2016; Feng et al. 2017; Corsi et al. 2023; T. Wang et al. 2022) (Supplementary Table S1). We filtered the initial dataset to obtain a thoroughly cleaned and curated dataset of 141 isolates (including the reference IOWA-ATCC genome), ensuring a robust foundation for reliable genomic analyses. Consistent with earlier findings (Corsi et al. 2023; T. Wang et al. 2022; Baptista et al. 2022), the overall genetic variability was modest, as just 28,047 biallelic high-quality SNPs were identified across the 141 genomes analysed. Phylogenetic analyses using both orthologous genes and SNP data provided robust evidence for the presence of two distinct lineages (Figure 1). One lineage (China/Egypt Lineage) consisted of all Chinese (IId) and Egyptian (IIa and IId) isolates, which was well distinct from the second lineage where all European (IIa and IId) and USA (IIa) isolates are grouped ("Western" lineage; see Fig. 1).

Considering that recombination has been, and still is, a fundamental driver of the genomic evolution of *C. parvum* (Corsi et al. 2023; T. Wang et al. 2022) and other *Cryptosporidium* species (Nader et al. 2019; Huang et al., 2023), we then focused on tracing these events. We described two clear events at chromosomes 1 and 4 (Figures or Table), involving isolates from different populations and hosts, with representatives of population 1 being the putative minor parents (i.e., the donors). Intriguingly, the very same event on chromosome 4 has already reported on a more limited dataset (Corsi et al 2023), while we provided extensive support for a recombination event on chromosome 1 that differed from that reported in Corsi et al. (2023). Additional evidence for the existence of mosaic genomes was obtained from network and admixture analyses of single chromosomes (Supplementary data and Figure). We observed SNP distribution patterns compatible with genetic exchanges, although the precise reconstruction of the events could not be achieved.

A deeper focus on the population structure showed that the “Western” lineage can be divided in two subgroups, namely population 3, a monophyletic group formed by all the USA and most of the European isolates, and population 2, a paraphyletic group that included the remaining European isolates.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 a slowed decay in LD in population 3 compared to population 2.

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 extra-European isolates, it seems reasonable to hypothesise that this coexistence of IIa and IId lineages in Europe could date back to ancient introductions of this parasite 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 (R. 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 all the USA isolates (from nine different States) form a monophyletic clade, indicating a single event of introduction from Europe, likely from the UK (Figure 1), and a subsequent expansion in the United States. Historical data ("The Introduction of Cattle into Colonial North America" 1942; Ficek 2019) and studies investigating the ancestry of New World cattles (McTavish et al. 2013; Delsol et al. 2023), suggest that this event should be relatively recent, as most of the import of livestock, particularly cattle, into the Americas occurred from the XVII to the XIX century by Portuguese and Spanish colonists, and during the Victorian Age by Britains (Ficek 2019; McTavish et al. 2013)).

Our results are consistent with *gp60* molecular typing data that identified only IIa subtypes in USA isolates (Jann et al. 2022). A parallel could be drawn with the recent emergence and rapid predominance of the *C. hominis* IfA12G1R5 subtype in the USA (Huang 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), while in the case of the *C. parvum* population 3, the data supports 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 (Figure 2), including all 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 foodborne outbreaks. While admittedly speculative, the most interesting result comes from a combination of population statistics and phylogeny-based statistical tests on gene sequences, which allowed to identify 16 candidate proteins under positive selection only in population 3. Interestingly, 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 fertilisation and transmission (Straschil et al. 2010). While 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 selective advantage. Other non-mutually exclusive explanations should be explored, including variation in 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.

Conclusions

Our study provides new insights into the epidemiology and evolution of *C. parvum*, with the description of a phylogenetic group that we indicated as the Western lineage. Within this Westen lineage, we observed the presence of two sympatric populations in Europe, and demonstrated that one has recently expanded to become predominant in young ruminants and humans, and then, likely from the UK, reached and spread into the USA. All isolates of the virulent and hyper transmissible IIaA15G2R1 subtype and all outbreak strains belonged to this recently expanding population, suggesting a selective advantage. Investigation of the genes under selective pressure identified a number of candidates with potential roles in the interaction with the host. Overall, our findings allow us to propose a scenario for the evolution of *C. parvum* and to pose focused questions for future research on this parasite.

Acknowledgements

We would like to 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 would like to thank Elke Radam (Robert Koch Institute, Germany) for excellent technical assistance.

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. Authors read and approved the final manuscript.

Funding information

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 773830: One Health European Joint Programme, PARADISE project (https://onehealthejp.eu/jrp-paradise/). TE and RRF were funded by the Development Fund Agricultural and Forestry MAKERA, Ministry of Agriculture and Forestry of Finland (grant number 435/03.01.02/2018).

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare that there are no conflicts of interest.

References

Alves, Margarida, Lihua Xiao, Irshad Sulaiman, Altaf A. Lal, Olga Matos, and Francisco Antunes. 2003. "Subgenotype Analysis of Cryptosporidium Isolates from Humans, Cattle, and Zoo Ruminants in Portugal." *Journal of Clinical Microbiology* 41 (6): 2744–47.

Baptista, Rodrigo P., Yiran Li, Adam Sateriale, Mandy J. Sanders, Karen L. Brooks, Alan Tracey, Brendan R. E. Ansell, et al. 2022. "Long-Read Assembly and Comparative Evidence-Based Reanalysis of Genome Sequences Reveal Expanded Transporter Repertoire and Duplication of Entire Chromosome Ends Including Subtelomeric Regions." *Genome Research* 32 (1): 203–13.

Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2021. "Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3." *eLife* 10 (May). https://doi.org/10.7554/eLife.65088.

Beja-Pereira, Albano, David Caramelli, Carles Lalueza-Fox, Cristiano Vernesi, Nuno Ferrand, Antonella Casoli, Felix Goyache, et al. 2006. "The Origin of European Cattle: Evidence from Modern and Ancient DNA." *Proceedings of the National Academy of Sciences of the United States of America* 103 (21): 8113–18.

Bhalchandra, Seema, Daviel Cardenas, and Honorine D. Ward. 2018. "Recent Breakthroughs and Ongoing Limitations in Research." *F1000Research* 7 (September). https://doi.org/10.12688/f1000research.15333.1.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20.

Cacciò, S. M., and R. M. Chalmers. 2016. "Human Cryptosporidiosis in Europe." *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 22 (6): 471–80.

Chalmers, Rachel M., Gregorio Pérez-Cordón, Simone M. Cacció, Christian Klotz, Lucy J. Robertson, and participants of the Cryptosporidium genotyping workshop (EURO-FBP). 2018. "Cryptosporidium Genotyping in Europe: The Current Status and Processes for a Harmonised Multi-Locus Genotyping Scheme." *Experimental Parasitology* 191 (August): 25–30.

Chalmers, R. M., and S. Cacciò. 2016. "Towards a Consensus on Genotyping Schemes for Surveillance and Outbreak Investigations of Cryptosporidium, Berlin, June 2016." *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 21 (37). https://doi.org/10.2807/1560-7917.ES.2016.21.37.30338.

Chavez, Miguel A., and A. Clinton White Jr. 2018. "Novel Treatment Strategies and Drugs in Development for Cryptosporidiosis." *Expert Review of Anti-Infective Therapy* 16 (8): 655–61.

Chessa, Bernardo, Filipe Pereira, Frederick Arnaud, Antonio Amorim, Félix Goyache, Ingrid Mainland, Rowland R. Kao, et al. 2009. "Revealing the History of Sheep Domestication Using Retrovirus Integrations." *Science* 324 (5926): 532–36.

Corsi, Giulia I., Swapnil Tichkule, Anna Rosa Sannella, Paolo Vatta, Francesco Asnicar, Nicola Segata, Aaron R. Jex, Cock van Oosterhout, and Simone M. Cacciò. 2023. "Recent Genetic Exchanges and Admixture Shape the Genome and Population Structure of the Zoonotic Pathogen Cryptosporidium Parvum." *Molecular Ecology* 32 (10): 2633–45.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. "The Variant Call Format and VCFtools." *Bioinformatics* 27 (15): 2156–58.

Darriba, Diego, David Posada, Alexey M. Kozlov, Alexandros Stamatakis, Benoit Morel, and Tomas Flouri. 2020. "ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models." *Molecular Biology and Evolution* 37 (1): 291–94.

Delsol, Nicolas, Brian J. Stucky, Jessica A. Oswald, Charles R. Cobb, Kitty F. Emery, and Robert

Guralnick. 2023. "Ancient DNA Confirms Diverse Origins of Early Post-Columbian Cattle in the Americas." *Scientific Reports* 13 (1): 1–12.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5): 491–98.

Dumaine, Jennifer E., Adam Sateriale, Alexis R. Gibson, Amita G. Reddy, Jodi A. Gullicksrud, Emma N. Hunter, Joseph T. Clark, and Boris Striepen. 2021. "The Enteric Pathogen Exports Proteins into the Cytosol of the Infected Host Cell." *eLife* 10 (December). https://doi.org/10.7554/eLife.70451.

Edgar, Robert C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32 (5): 1792–97.

Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.

Feng, Yaoyu, Na Li, Dawn M. Roellig, Alyssa Kelley, Guangyuan Liu, Said Amer, Kevin Tang, Longxian Zhang, and Lihua Xiao. 2017. "Comparative Genomic Analysis of the IId Subtype Family of Cryptosporidium Parvum." *International Journal for Parasitology* 47 (5): 281–90.

Feng, Yaoyu, Una M. Ryan, and Lihua Xiao. 2018. "Genetic Diversity and Population Structure of Cryptosporidium." *Trends in Parasitology* 34 (11): 997–1011.

Ficek, Rosa E. 2019. "Cattle, Capital, Colonization." *Current Anthropology* 60 (S20): S260–71.

Garcia-R, Juan C., and David T. S. Hayman. 2016. "Origin of a Major Infectious Disease in Vertebrates: The Timing of Cryptosporidium Evolution and Its Hosts." *Parasitology* 143 (13): 1683–90.

Gruber-Vodicka, Harald R., Brandon K. B. Seah, and Elmar Pruesse. 2020. "phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes." *mSystems* 5 (5). https://doi.org/10.1128/mSystems.00920-20.

Guo, Yaqiong, Una Ryan, Yaoyu Feng, and Lihua Xiao. 2021. "Association of Common Zoonotic Pathogens With Concentrated Animal Feeding Operations." *Frontiers in Microbiology* 12: 810142.

Hadfield, Stephen J., Justin A. Pachebat, Martin T. Swain, Guy Robinson, Simon Js Cameron, Jenna Alexander, Matthew J. Hegarty, Kristin Elwin, and Rachel M. Chalmers. 2015. "Generation of Whole Genome Sequences of New Cryptosporidium Hominis and Cryptosporidium Parvum Isolates Directly from Stool Samples." *BMC Genomics* 16 (1): 650.

Hemstrom, William, and Melissa Jones. 2023. "snpR: User Friendly Population Genomics for SNP Data Sets with Categorical Metadata." *Molecular Ecology Resources* 23 (4): 962–73.

Hijjawi, Nawal, Alizera Zahedi, Mohammed Al-Falah, and Una Ryan. 2022. "A Review of the Molecular Epidemiology of Cryptosporidium Spp. and Giardia Duodenalis in the Middle East and North Africa (MENA) Region." *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 98 (March): 105212.

Huson, Daniel H., and David Bryant. 2006. "Application of Phylogenetic Networks in Evolutionary Studies." *Molecular Biology and Evolution* 23 (2): 254–67.

Innes, Elisabeth A., Rachel M. Chalmers, Beth Wells, and Mattie C. Pawlowic. 2020. "A One Health Approach to Tackle Cryptosporidiosis." *Trends in Parasitology* 36 (3): 290–303.

Jann, Higor Wilson, Mauro Jorge Cabral-Castro, João Victor Barreto Costa, Alba Cristina Miranda de Barros Alencar, José Mauro Peralta, and Regina Helena Saramago Peralta. 2022. "Prevalence of Human Cryptosporidiosis in the Americas: Systematic Review and Meta-Analysis." *Revista Do Instituto de Medicina Tropical de Sao Paulo* 64. https://doi.org/10.1590/S1678-9946202264070.

Khalil, Ibrahim A., Christopher Troeger, Puja C. Rao, Brigette F. Blacker, Alexandria Brown, Thomas G. Brewer, Danny V. Colombara, 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." *The Lancet. Global Health* 6 (7): e758–68.

Khan, Shahbaz M., and William H. Witola. 2023. “Past, Current, and Potential Treatments for Cryptosporidiosis in Humans and Farm Animals: A Comprehensive Review.” *Frontiers in Cellular and Infection Microbiology* 13 (January): 1115522.

Kotloff, Karen L., James P. Nataro, William C. Blackwelder, Dilruba Nasrin, Tamer H. Farag, Sandra Panchalingam, Yukun Wu, 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.” *The Lancet* 382 (9888): 209–22.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Long, Shaojun, Bryan Anthony, Lisa L. Drewry, and L. David Sibley. 2017. “A Conserved Ankyrin Repeat-Containing Protein Regulates Conoid Stability, Motility and Cell Invasion in Toxoplasma Gondii.” *Nature Communications* 8 (1): 1–14.

Manske, Magnus, Olivo Miotto, Susana Campino, Sarah Auburn, Jacob Almagro-Garcia, Gareth Maslen, Jack O’Brien, et al. 2012. “Analysis of Plasmodium Falciparum Diversity in Natural Infections by Deep Sequencing.” *Nature* 487 (7407): 375–79.

Martin, Darren P., Ben Murrell, Michael Golden, Arjun Khoosal, and Brejnev Muhire. 2015. “RDP4: Detection and Analysis of Recombination Patterns in Virus Genomes.” *Virus Evolution* 1 (1): vev003.

Mathur, Varsha, Kevin C. Wakeman, and Patrick J. Keeling. 2021. “Parallel Functional Reduction in the Mitochondria of Apicomplexan Parasites.” *Current Biology: CB* 31 (13): 2920–28.e4.

McKerr, Caoimhe, Sarah J. O’Brien, Rachel M. Chalmers, Roberto Vivancos, and Robert M. Christley. 2018. “Exposures Associated with Infection with Cryptosporidium in Industrialised Countries: A Systematic Review Protocol.” *Systematic Reviews* 7 (1): 70.

McTavish, Emily Jane, Jared E. Decker, Robert D. Schnabel, Jeremy F. Taylor, and David M. Hillis. 2013. “New World Cattle Show Ancestry from Multiple Independent Domestication Events.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (15): E1398–1406.

Murrell, Ben, Steven Weaver, Martin D. Smith, Joel O. Wertheim, Sasha Murrell, Anthony Aylward, Kemal Eren, et al. 2015. “Gene-Wide Identification of Episodic Selection.” *Molecular Biology and Evolution* 32 (5): 1365–71.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. “Inference of Population Structure Using Multilocus Genotype Data.” *Genetics* 155 (2): 945–59.

Puiu, Daniela, Shinichiro Enomoto, Gregory A. Buck, Mitchell S. Abrahamsen, and Jessica C. Kissinger. 2004. “CryptoDB: The Cryptosporidium Genome Resource.” *Nucleic Acids Research* 32 (Database issue): D329–31.

Rahman, Sajid Ur, Rongsheng Mi, Shasha Zhou, Haiyan Gong, Munib Ullah, Yan Huang, Xiangan Han, and Zhaoguo Chen. 2022. “Advances in Therapeutic and Vaccine Targets for Cryptosporidium: Challenges and Possible Mitigation Strategies.” *Acta Tropica* 226 (February): 106273.

Ryan, Una M., Yaoyu Feng, Ronald Fayer, and Lihua Xiao. 2021. “Taxonomy and Molecular Epidemiology of Cryptosporidium and Giardia - a 50 Year Perspective (1971-2021).” *International Journal for Parasitology* 51 (13-14): 1099–1119.

Ryan, Una, Lihua Xiao, Carolyn Read, Ling Zhou, Altaf A. Lal, and Ivan Pavlasek. 2003. “Identification of Novel Cryptosporidium Genotypes from the Czech Republic.” *Applied and Environmental Microbiology* 69 (7): 4302–7.

Ryan, Una, Alireza Zahedi, Yaoyu Feng, and Lihua Xiao. 2021. “An Update on Zoonotic Species and Genotypes in Humans.” *Animals : An Open Access Journal from MDPI* 11 (11). https://doi.org/10.3390/ani11113307.

Schaffner, Stephen F., Aimee R. Taylor, Wesley Wong, Dyann F. Wirth, and Daniel E. Neafsey. 2018. “hmmIBD: Software to Infer Pairwise Identity by Descent between Haploid Genotypes.” *Malaria Journal* 17 (1): 196.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.

Stanke, Mario, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. 2006. “AUGUSTUS: Ab Initio Prediction of Alternative Transcripts.” *Nucleic Acids Research* 34 (Web Server issue): W435–39.

Straschil, Ursula, Arthur M. Talman, David J. P. Ferguson, Karen A. Bunting, Zhengyao Xu, Elizabeth Bailes, Robert E. Sinden, et al. 2010. “The Armadillo Repeat Protein PF16 Is Essential for Flagellar Structure and Function in Plasmodium Male Gametes.” *PloS One* 5 (9): e12901.

“The Introduction of Cattle into Colonial North America.” 1942. *Journal of Dairy Science* 25 (2): 129–54.

Troell, Karin, Björn Hallström, Anna-Maria Divne, Cecilia Alsmark, Romanico Arrighi, Mikael Huss, Jessica Beser, and Stefan Bertilsson. 2016. “Cryptosporidium as a Testbed for Single Cell Genome Characterization of Unicellular Eukaryotes.” *BMC Genomics* 17 (June): 471.

Tůmová, Lenka, Jana Ježková, Jitka Prediger, Nikola Holubová, Bohumil Sak, Roman Konečný, Dana Květoňová, et al. 2023. “Cryptosporidium Mortiferum N. Sp. (Apicomplexa: Cryptosporidiidae), the Species Causing Lethal Cryptosporidiosis in Eurasian Red Squirrels (Sciurus Vulgaris).” *Parasites & Vectors* 16 (1): 1–21.

Van der Auwera, Geraldine A., and Brian D. O’Connor. 2020. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O’Reilly Media.

Wang, Rongjun, Longxian Zhang, Charlotte Axén, Camilla Bjorkman, Fuchun Jian, Said Amer, Aiqin Liu, et al. 2014. “Cryptosporidium Parvum IId Family: Clonal Population and Dispersal from Western Asia to Other Geographical Regions.” *Scientific Reports* 4 (1): 1–5.

Wang, Tianpeng, Yaqiong Guo, Dawn M. Roellig, Na Li, Mónica Santín, Jason Lombard, Martin Kváč, et al. 2022. “Sympatric Recombination in Zoonotic Cryptosporidium Leads to Emergence of Populations with Modified Host Preference.” *Molecular Biology and Evolution* 39 (7). https://doi.org/10.1093/molbev/msac150.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wick, Ryan R., Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. 2017. “Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads.” *PLoS Computational Biology* 13 (6): e1005595.

Zahedi, Alireza, and Una Ryan. 2020. “Cryptosporidium - An Update with an Emphasis on Foodborne and Waterborne Transmission.” *Research in Veterinary Science* 132 (October): 500–512.

Zhang, Chi, Shan-Shan Dong, Jun-Yang Xu, Wei-Ming He, and Tie-Lin Yang. 2019. “PopLDdecay: A Fast and Effective Tool for Linkage Disequilibrium Decay Analysis Based on Variant Call Format Files.” *Bioinformatics* 35 (10): 1786–88.

Zhou, Hua, David Alexander, and Kenneth Lange. 2011. “A Quasi-Newton Acceleration for High-Dimensional Optimization Algorithms.” *Statistics and Computing* 21 (2): 261–73.

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. *Infection Genetics and Evolution*. 98:105212. doi: 10.1016/j.meegid.2022.105212.

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: https:// doi. org/ 10. 1186/ s12864- 019- 5788- 9