Article - Discoveries

Title

Global population genomics of two subspecies of *Cryptosporidium hominis* during 500 years of evolution

Author names

Swapnil Tichkule[1,2,3*], Simone M. Cacciò[4*], Guy Robinson[5,6], Rachel M. Chalmers[5,6], Ivo Mueller[1,3], Samantha J. Emery-Corbin[1], Daniel Eibach[7,8], Kevin M. Tyler[9,10], Cock van Oosterhout[11*] and Aaron R. Jex[1,12*]

Affiliation

[1]Population Health and Immunity, Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC, Australia;

[2]Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, VIC, Australia;

[3]Department of Medical Biology, University of Melbourne, Melbourne, Australia;

[4]Department of Infectious Disease, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161, Rome, Italy;

[5]Cryptosporidium Reference Unit, Public Health Wales Microbiology, Singleton Hospital,
Swansea, UK;

[6]Swansea University Medical School, Swansea, UK;

[7]Department of Infectious Disease Epidemiology, Bernhard Nocht Institute for Tropical Medicine Hamburg, Bernhard-Nocht-Strasse 74, 20359 Hamburg, Germany;

[8]German Center for Infection Research (DZIF), Hamburg-Lübeck-Borstel-Riems, Germany;

[9]Biomedical Research Centre, Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, UK;

[10]Center of Excellence for Bionanoscience Research, King Abdul Aziz University, Jeddah, Saudi Arabia

[11]School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, UK;

[12]Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Melbourne, VIC, Australia

*** Corresponding authors:**

Swapnil Tichkule

Postal address: Population Health & Immunity, Walter and Eliza Hall Institute of Medical Research,
1G, Royal Parade, Parkville VIC 3052
Phone: +61 404651636
e-mail: tichkule.s@wehi.edu.au

Simone M. Cacciò
Postal address: Department of Infectious Diseases, Istituto Superiore di Sanità, Viale Regina Elena
299, Rome 00161 Italy
Phone: +39 06 4990 2304
e-mail: simone.caccio@iss.it

Cock van Oosterhout
Postal address: School of Environmental Sciences, University of East Anglia, Norwich Research
Park, Norwich NR4 7TJ, United Kingdom
Phone: +44 (0)1603 592921
e-mail: c.van-oosterhout@uea.ac.uk

Aaron Jex
Postal address: Population Health & Immunity Division, Walter & Eliza Hall Institute, 1G Royal
Parade, Parkville, VIC 3052
Phone: +61 3 9345 2291
e-mail: jex.a@wehi.edu.au

Abstract

Cryptosporidiosis is a major global health problem and a primary cause of diarrhoea, particularly in young children in low- and middle-income countries (LMICs). The zoonotic *Cryptosporidium parvum* and anthroponotic *C. hominis* cause most human infections. Here, we present a comprehensive whole-genome study of *C. hominis,* comprising 114 isolates from 16 countries within five continents. We detect two lineages with distinct biology and demography, which diverged circa 500 years ago. We consider these lineages two subspecies and propose the names *C. hominis hominis* and *C. hominis aquapotentis* (*gp60* subtype IbA10G2). In our study, *C. h. hominis* is almost exclusively represented by isolates from LMICs in Africa and Asia and appears to have undergone recent population contraction. In contrast, *C. h. aquapotentis* was found in high-income countries, mainly in Europe, North America and Oceania, and appears to be expanding. Notably, *C. h. aquapotentis* is associated with high rates of direct human-to-human transmission, which may explain its success in countries with well-developed environmental sanitation infrastructure. Intriguingly, we detected

genomic regions of introgression following secondary contact between the subspecies. This resulted in high diversity and divergence in genomic islands of putative virulence genes (GIPVs), including *muc5 (*CHUDEA2_430) and a hypothetical protein (CHUDEA6_5270). This diversity is maintained by balancing selection, suggesting a coevolutionary arms race with the host. Lastly, we find that recent gene flow from *C. h. aquapotentis* to *C. h. hominis*, likely associated with increased human migration, may be driving evolution of more virulent *C. hominis* variants.

Introduction

Cryptosporidiosis is a leading cause of diarrhoea in children under five globally (Kotloff, et al. 2013; Khalil, et al. 2018), resulting in an estimated 48,000 deaths annually. Among parasitic diseases, it is second only to malaria in global health burden, with an overall impact of ~12.8 million Disability Adjusted Life-Years (DALYs) (Khalil, et al. 2018). Human cryptosporidiosis is primarily caused by *Cryptosporidium parvum*, a zoonoses common in young ruminants (Ryan, et al. 2016; Santin 2020), and *C. hominis*, which is anthroponotic and the more prevalent species globally (Razakandrainibe, et al. 2018). The disease burden is overwhelmingly skewed to low and middle-income countries (LMICs) (Yang, et al. 2021), particularly in sub-Saharan Africa (Khalil, et al. 2018), where *C. hominis* predominates. However, *Cryptosporidium* remains a significant public health problem in wealthy countries through large water or foodborne outbreaks and direct transmission in day-care facilities, hospitals and other institutions (Craun, et al. 1998; Putignani and Menichella 2010). Presently, there are no vaccines or effective drugs to treat the infection. Hence, control depends on prevention of infection, which is driven by a strong understanding of the parasite's epidemiology.

The global epidemiology of cryptosporidiosis varies by geographic region, socioeconomic status and a range of risk factors (Nichols, et al. 2014; Yang, et al. 2021). Understanding of this epidemiology is underpinned by molecular typing, based mainly on the highly polymorphic *gp60* gene (Feng, et al. 2018). This work has identified numerous genetic variants within each species and indicated a complex population genetic structuring. In *C. parvum*, population structure varies globally from clonal to epidemic to panmictic, likely due to varying ecological factors (Morrison, et al. 2008; Herges, et al. 2012; Wang, et al. 2014). Genetic variants found exclusively in humans point to an anthroponotic *C. parvum* lineage (IIc) (King, et al. 2017; King, et al. 2019), with a recent genomic study recognising two subspecies, the zoonotic *C. parvum parvum* and the anthroponotic *C. parvum anthroponosum (*Nader, et al. 2019). Interestingly, this study found that both subspecies occasionally still hybridise and exchange genetic variation. These exchanges overlapped with similar genomic regions undergoing genetic introgression between *C. hominis* and *C. parvum anthroponosum*, indicating candidate sites underpinning adaptation to human-specific infection (Nader, et al. 2019). The signature of introgression can be identified by comparing the DNA sequence (dis)similarity between two or more haplotypes. Briefly, introgressed regions are characterised by high nucleotide similarity, resulting from a relatively recent coalescence of the introgressed sequence. By comparing the sequence variation of three or more haplotypes, the directionality of genetic exchange can be inferred. In addition, the genetic divergence (i.e., number of nucleotide substitutions) can be used to estimate how long ago the sequence was exchanged between the ancestors of the haplotypes. These are the principles of software such as Hybrid-Check (Ward & van Oosterhout 2016) that enable the study of genetic introgression. *Cryptosporidium hominis* appears to have a largely clonal population structure, dominated by specific variants in different regions (Yang, et al. 2021). The IbA10G2 *gp60* subtype, although found in LMICs (Jex and Gasser 2010), is the dominant variant in high-income countries and accounts for up to ~45% of all *gp60*-typed human infections (Jex and Gasser 2010). This subtype is linked to most major waterborne outbreaks in wealthy countries for which genetic typing is available (Zhou, et al. 2003; Chalmers, et al. 2010; Widerström, et al. 2014; Segura, et al. 2015; Efstratiou, et al. 2017). The IbA10G2 subtype also appears more readily capable of direct human-to-human transmission (McKerr, et al. 2021) and may cause more severe disease (Cama, et al. 2008).

Studies of *C. hominis* molecular epidemiology pose several essential questions that have major implications for global control of cryptosporidiosis. Specifically: (1) what is the IbA10G2 subtype, and is this *gp60*-defined subtype reflective of a phylogenomically divergent *C. hominis* lineage that predominates in wealthy countries; (2) if so, does this lineage undertake reduced levels of genetic recombination with global *C. hominis* populations; (3) and can signatures within the genome sequences of IbA10G2 typed isolates identify its taxonomic status, reveal the factors underpinning its putatively increased virulence and its dominance in wealthy countries, and identify its influence on global parasite population structure? To address these questions, we performed a global study of 114 *C. hominis* genome sequences, comparing the IbA10G2 subtype to published genome sequences representing locally acquired infections from 16 countries across five continents. The insights gained from these analyses are particularly relevant for public health; the IbA10G2 subtype has already been identified as an emerging host-adapted, likely more virulent and transmissible population, making it a threat to human health in high-income countries (Li, et al. 2013; Feng, et al. 2014; Cacciò and Chalmers 2016).

Results

Genomic evidence of population sub-structuring at the continental level

All 114 *C. hominis* isolates included in this study were confirmed as single variant infections (estimated MOI=1 and *Fws* > 0.95) (see Supplementary Table 2). We identified 5,618 biallelic SNPs among these samples and used these to explore the population structure of *C. hominis*. Our analyses identified two major clusters (Fig. 1a), separating European, North American and Oceanian samples from Asian and African samples. We also saw minor clustering separating African from Asian samples. STRUCTURE analysis provided further support for these findings (Fig. 1b). Overall, ~94% of isolates are clustered geographically. Exceptions included a small number of infections (e.g., UKH30 (from UK) acquired during international travel. STRUCTURE analysis also identified a fourth cluster, including three African and one European isolate, with unique population ancestry (cluster 4 in Fig. 1b).

Genomic evidence of population diversification

Maximum Likelihood (ML) analysis (Fig. 1c) identified two major clades, one corresponding to Asia and Africa (clade 1) and the other to Europe, North America and Oceania (clade 2). All isolates within clade 2 had *gp60* subtype IbA10G2, and no IbA10G2 isolates clustered with clade 1. The two clades were estimated to have diverged 488 (84 – 2,199; 95% HPD) years ago. This divergence is supported by a SplitTree network (Fig. 1d) and a DensiTree (Fig. 1e) analyses. Considering the stability of these two genomically distinct lineages and evidence below of their reproductive isolation, we propose their recognition as separate subspecies. We propose clade 1, which comprises infections observed in low- to middle-income countries, mostly from Africa and Asia, be recognised as *C. hominis hominis*, referring to the fact that this subspecies represents the majority population. We propose clade 2 (IbA10G2 subtype) includes isolates from high-income countries, namely Europe, North America and Oceania, be named *C. hominis aquapotentis (*strong water), as it predominates in countries with longstanding high sanitation and water quality indices.

Demographic histories

To understand the demographic histories and estimate the change in the effective population size (*Ne*) through time, we constructed a Bayesian Skyline Plot (BSP) (Drummond, et al. 2005) for *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2 IbA10G2). Parasite isolates from low-income countries (*C. h. hominis,* clade 1) experienced a marked population contraction recently from *Ne*=~5000 to *Ne*=~100 (Fig. 2a), which is supported by a higher proportion of positive Tajima's D values (Fig. 2b). In contrast, the isolates from high-income countries (*C. h. aquapotentis* (clade 2), *gp60* subtype IbA10G2) had a stable effective population size (*Ne*=~1000) and a higher proportion of negative Tajima's D values (Fig. 2c). The distribution of Tajima's D in *C. h. aquapotentis* is significantly smaller than zero (one-sided t-test: t = -28.883, df = 681, p-value < 2.2e-16), which is consistent with recent population expansion. BSP analysis of *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) suggests a stable *Ne*, whereas the overall negative value of Tajima's D analysis is consistent with a recent population size expansion or selective sweep. BSP analyses are based on coalescent theory, whereas Tajima's D compares the constitution of polymorphisms, i.e., mean number of pairwise differences and the number of segregating sites. The latter is more sensitive to recent demographic events. Altogether, our analyses thus imply that after a relatively stable *Ne*, the population has started to expand only very recently, or that it has been affected by a recent selective sweep. Simulation-based projections of the evolution of the overall *C. hominis* population finds it is being shaped by 'recent gene flow', likely indicating recent secondary contact (Fig. 2c-d) between the two subspecies ~165 (24 – 647; 5-95% CI) years ago.

We found evidence of two selective sweeps on chromosome 6 in *C. h. hominis* (clade 1) based on composite likelihood ratio (CLR) statistic using SweeD (Pavlidis, et al. 2013) (Supplementary Fig. 1). However, we do not see any other hallmarks of a selective sweep in these regions based on π or Tajima's D (see Supplementary Fig. 2). Possibly, the selective sweeps were incomplete or occurred in the distant past, eroding their genomic signatures. Nucleotide diversity and Tajima's D are simple statistical approaches to identify selective sweep, but they are also affected by population size changes. In contrast, SweeD analyses site frequency spectra (SFS) and uses complex statistical approaches such as likelihood-based methods to identify selective sweeps in whole genome data. However, the likelihood of selective sweep identified in our analysis is low which might be the reason that Nucleotide diversity and Tajima's D methods did not identify any statistically significant sweeps. Nevertheless, this may have contributed to the recent decline in effective population size of *C. h. hominis* (clade 1) *(*Fig. 2a); genetic variation could have been lost during the selective sweep not only in the affected chromosomal regions, but throughout the genome in this largely clonally reproducing organisms, as the selectively favoured variant replaced other existing variants.

Linkage, recombination, introgression and gene flow between subspecies

Distinct patterns of decay of linkage disequilibria

We performed independent linkage analyses for each subspecies to infer the recombination rate within parasite populations from low- and high-income countries (Fig. 3a). *Cryptosporidium h. hominis* (clade 1) had more rapid linkage disequilibrium (LD) decay than *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2), consistent with genetic exchanges through gene flow and recombination. In contrast, the strong LD in *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) supports our hypothesis of a recent population expansion (see below).

Recombination and regions of secondary contact

Using RDP4 (Martin, et al. 2015), we identified significant recombination events between the two clades in chromosome 1 (event 1), 2 (event 2) and 6 (event 3 and 4) (Fig; 3b and Supplementary

Table 3). This indicated secondary contacts between *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2), which resulted in rare genetic exchanges between these otherwise diverged clades. We also undertook additional analyses of the highly admixed European isolate (UK_UKH4 of *gp60* subtype IaA14R3) that showed unique ancestry (Fig 1B) and clustered with low-income countries (*C. h. hominis* (clade 1)) (see Supplementary Text).

Signature of introgression and gene flow between the clades

We analysed the recombination events in more detail to better understand the implications of genetic introgression between the two subspecies. Determining the signature of genetic introgression is crucial as these regions can also be responsible for increasing genetic diversity and providing novel substrate for natural selection in host-parasite coevolution (Van Oosterhout 2021). We used HybridCheck (Ward and van Oosterhout 2016) to perform introgression analyses for recombinant events 2 – 4 (event 1 was excluded due to a missing parental sequence). We randomly selected a triplet (recombinant, minor parent and major parental lines) from the RDP output (Supplementary Table 3), which revealed a clear signature of introgression (Fig. 3 c-d), also supported by an ABBA-BABA test (Fig. 3e). Additionally, we calculated the pairwise r^2 between SNPs within chromosomes of *C. h. hominis* (clade 1) to assess linkage among SNPs in the introgressed regions (Fig. 3f-g). Large blocks of high LD that encompass the introgressed regions suggested each had been exchanged as a single event between *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) ~165 (24 – 647; 5-95% CI) years ago.

Could recent introgression increase virulence?

Our analyses suggest *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) have diverged and largely reproductively isolated for ~500 years. Noting that the latter subspecies has come to dominate infections within high-income countries (Guo, et al. 2015), is more virulent (Cama, et al. 2008), appears better able to transmit through direct human-to-human contact (McKerr, et al. 2021) and possibly at a lower infectious dose (Segura, et al. 2015), *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) may owe its success to being better adapted to human infection. If this is the case, it is possible that recent introgression between these two subspecies could select for more virulent and transmissible *C. hominis* subspecies in low-income countries, particularly noting the recent genetic bottlenecking we have observed in *C. h. hominis* (clade 1) within the last decades.

Identification of potential virulence genes

To explore this, we first predicted candidate virulence genes in *C. hominis* likely to be involved in the host-parasite interaction and engaged in a coevolutionary arms race with the host. Broadly, such genes are under continuous adaptation while interacting with hosts (Van Oosterhout 2021). During coevolution, evolutionary forces act on virulence genes to create genetic variation through mutation, recombination and gene flow and this variation is moulded by natural selection (and genetic drift). To identify putative virulence genes in *C. hominis*, we selected the top 5% most highly polymorphic genes based on nucleotide diversity (π). These were filtered for the top 25% of genes under balancing selection, based ranked Tajima's D. Finally, these genes were filtered by selecting the top 50% of genes with the highest proportion of non-synonymous mutations, based on ranked Ka/Ks ratio. Using this approach, we identified 24 highly polymorphic genes (Supplementary Table 5) that are rapidly mutating at the protein level and under selective pressure. These genes were significantly more polymorphic (two-sided t-test: t = -2.9062, df = 23, p-value = 0.007957), under stronger balancing selection (two-sided t-test: t = -10.393, df = 23, p-value = 2.533e-10) and positive selection (two-sided t-test: t = -2.4736, df = 23, p-value = 0.021) compared to all remaining polymorphic genes. Moreover, these genes are enriched in recombination regions ($\chi 2$= 225.04, df = 1, p-value = 0.00049), showing that gene flow within *C. hominis* is inclined towards elevated levels of genetic variation. Overall, these genes are enriched for extracellular ($\chi 2$= 13.608, df = 1, p-value = 0.00349) and signal peptide proteins ($\chi 2$= 6.69, df = 1, p-value = 0.015), which is consistent with their potential relevance for host-parasite interaction. Together, these results provided the evidence of genomic islands of putative virulence genes (GIPVs: Fig. 4) that are likely to be in a coevolutionary arms race with the host, undergoing frequent recombination and accumulating beneficial polymorphisms that are under selection, and that this increased population diversity.

Subspecific divergence of putative virulence genes

We investigated whether any virulence genes predicted above were highly diverged between *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) and under diversifying selection between the two subspecies. To do this, we calculated absolute divergence (Dxy), representing the average proportion of differences between all pairs of sequences between *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2), revealing differential gene flow during their reproductive isolation (Fig. 4a). We then calculated the correlation between diversity (π) and divergence (Dxy) for each gene (Fig. 4b). This approach identified four clear outlier genes, which were the most divergent, diverse, and rapidly mutating at the protein level. These genes encoded three mucins (CHUDEA2_430, CHUDEA2_440 and CHUDEA2_450) arrayed in a cluster on chromosome 2, and a hypothetical protein (CHUDEA6_5270) found on chromosome 6 (Fig. 4b, Supplementary Table 5). The *gp60* gene ranked 6th in the list of virulence genes but does not sit as a clear outlier from the other genes identified in our assessment. Two of these genes, CHUDEA2_430 (LRT, p-value = 0.005) and CHUDEA6_5270 (LRT, p-value = 0.017), are under statistically significant diversifying selection between *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2).

Lastly, we looked at each codon position within each of the four outlier genes to detect codon specific diversifying selection, which might be overlooked in the overall gene. We detected episodic diversifying selection (LRT p-value < 0.05) at codon positions 92, 111 and 138 of CHUDEA2_430 and 97 and 105 of CHUDEA6_5270 using the Mixed Effects Model of Evolution (MEME) method implemented in Datamonkey (Pond and Frost 2005). These sites represent putative codons harbouring genetic polymorphisms that experience periods of strong diversifying selection. This strongly suggests CHUDEA2_430 and CHUDEA6_5270 are likely candidate virulence factors participating in a coevolutionary arms race and contributing to the divergence of *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2).

Recent introgression of putative virulence genes

Finally, we asked whether the recent recombination events in chromosomes 2 and 6 between *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) might include genes associated with increased virulence, indicating increasing virulence of *C. hominis* globally within the past decades, possibly linked to increased human migration. These introgressed regions had significantly elevated levels of nucleotide diversity (two-sided t-test: t = -3.0026, df = 27, p-value = 0.0057), divergence (two-sided t-test = -3.0043, df = 27, p-value = 0.0057) and balancing selection (two-sided t-test: t = -3.0125, df = 27, p-value = 0.0055). Interestingly, we observed a pattern of high diversity and divergence (Fig. 4a), particularly for genes within the recombinant blocks (Fig. 3b-e). The 38 (= top 1%) most divergent genes between *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) were enriched in recombinant regions ($\chi2$= 287.08, df = 1, p-value = 0.00049), with 37% present in these regions (Supplementary Table 6), compared to 0.9% of the other 1,645 divergent genes. These results suggest the genes undergoing frequent recombination accumulated beneficial polymorphisms that were maintained by balancing selection and that this increased population diversity.

Notably, the introgressed regions between *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) included CHUDEA2_430 (*muc5*) for event 2 (chromosome 2), CHUDEA6_1080 (*gp60*) for event 3 (chromosome 6) and CHUDEA6_5270 (a hypothetical protein) for event 4 (chromosome 6). The introgression at CHUDEA6_5270 is particularly intriguing as it sheds further light on the evolution of *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) and *C. h. hominis* (clade 1), revealing how recent recombination events could be driving the virulence evolution of *C. h. hominis* (clade 1). We identified two major CHUDEA6_5270 (hypothetical gene) haplotypes; Hap1 representing most *C. h. hominis* (clade 1) isolates, and Hap2 represents all *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) plus a small subset of *C. h. hominis* (clade 1) (Fig. 5a), as well as many haplotypes associated only with *C. h. hominis* (clade 1). No mutations were observed in Hap2, which is consistent with *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) having evolved recently, being an estimated 392 years (29 – 1699 years; 5-95% CI) old, assuming a mutation rate of $u=10^{-8}$ and 48h cell division time. Hap1 and Hap2 are diverged by 22 SNPs (Fig. 5b), which is unlikely to represent standing genetic variation, given this is 3-fold higher than the mean deviation among all other CHUDEA6_5270 haplotypes identified here. Instead, we propose CHUDEA6_5270-Hap2 might be an introgressed variant from a highly diverged, unsampled (sub)species that diverged from *C. hominis* around 2.36 million generations ago (assuming a mutation rate of $u=10^{-8}$), which is equal to 12,742 (8,901 – 17,497 years; 5-95% CI) years (assuming 48h/replication). The emergence of CHUDEA6_5270-Hap2 in some *C. h. hominis* (clade 1) isolates (i.e., the red section of Hap2 in Fig. 6a) strongly implies that *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) has introgressed (circa 400 years ago) into *C. h. hominis* (clade 1) leading to evolution under balancing selection (Fig. 5d).

We modelled the 3D protein structure of CHUDEA6_5270-Hap1 and Hap2, and compared these to similarly predicted 3D protein structures for the orthologous genes from *C. parvum parvum* and *C. parvum anthropanosum (*Fig. 5c). This modelling identified several conserved alpha-helices that appear to form a coiled-coil domain. CHUDEA6_5270 from *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) encodes mutations near the C-terminal end, resulting in a notable kink that deviates from the other structures. This structural variation overlaps with an increase in Tajima's D values toward the C-terminus of the protein, suggesting the region is under balancing selection (Fig. 5d).

A similar pattern is observed for CHUDEA2_430 (see Supplementary Fig. 6). However, given that both the subspecies are interspersed in the haplotype network, indicating against a specific directionality of introgression. This may indicate that CHUDEA_430 diverged with the divergence of the subspecies and continued to diversify. We were not able to generate a robust 3D structural model for CHUDEA2_430 or its *C. parvum* orthologs. We noted the gene encodes a large intrinsically disordered region (Supplementary Fig. 7) which is a consistent feature of mucin proteins (Carmicheal, et al. 2020), whose structural confirmation is influenced by post-translation glycosylation (Perez-Vilar and Hill 1999). Noting this, we did identify eight novel glycosylation sites in *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) CHUDEA2_430 haplotypes not found in *C. h. hominis* (clade 1) *(*Supplementary Fig 5). Whether these impact interaction with host proteins is not known, but this would be consistent with glycosylated proteins in other pathogens (Lin, et al. 2020).

Discussion

In this study, we examined the evolutionary genomics of a major human parasite, *C. hominis*, studying whole genome sequence data from 114 isolates from 16 countries across five continents. We posed three questions, and we are able to answers these as follows: (1) the *gp60*-defined subtype IbA10G2 is reflective of a phylogenomically divergent *C. hominis* lineage that predominates in wealthy countries; (2) this lineage has experienced significantly reduced levels of genetic recombination with other global *C. hominis* populations, and we could identify only four genetic exchanges between these otherwise diverged clades; (3) the distinct evolutionary trajectory of the IbA10G2 subtype, characterized by rapid population expansion, warrants a distinct taxonomic status as subspecies. We propose the name *C. hominis aquapotentis* (clade 2, *gp60* subtype IbA10G2) to reflect its adaptation to "strong water", i.e., high sanitation and water quality indices.

These two subspecies are estimated to have been reproductively isolated for approximately 488 (84 – 2,199) years, except for the more recent genetic exchange at four genetic loci. The reproductive isolation coincides with improvements to sanitation in Europe. It is possible that *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) evolved specialisations making it better suited to human infection allowing it to be more successful through direct transmission supported by a lower infectious dose. Such adaptations would have allowed *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) to become dominant in higher-income countries where sanitation has reduced the level of environmental transmission. Indeed, epidemiological investigations in the UK identified direct person-to-person transmission as a key pathway for *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) (McKerr, et al. 2021). We hypothesise this may have resulted in its rapid population expansion over the last ~500 years and (partial) reproductive isolation. In contrast, *C. h. hominis* (clade 1) experienced recent reduction in effective population size (*Ne*) within the past few decades. We detected two signatures of recent selective sweep in this subspecies, and we propose that this may have eroded some of the genetic variation, resulting in the marked drop in *Ne* = 5000 to *Ne* = 100 in the past. This would have resulted in significant genetic drift and random allele frequency changes, which could have increased the divergence between *C. h. hominis* (clade 1) and *C. hominis aquapotentis* (clade 2, *gp60* subtype IbA10G2) further.

Despite increased migration and international travelling in the past decades, our study suggests that there has been relatively little movement between continents for this parasite. This is in contrast to reports for *C. parvum.* Corsi, *et. al.* recently found a higher proportion of admixture and gene flow between *C. parvum* populations and no evidence of population structuring by geographic region (Corsi, et al. 2021). Despite the strong population sub-structuring in *C. hominis*, we found evidence of potential recombination and gene flow between the geographic populations and subspecies. We further investigated and identified the introgressed regions where we detected significant gene flow between the low- and high- income countries. Simulation-based analyses indicated this was most likely explained by 'recent geneflow' (circa 165 years ago). This would appear to be a secondary contact between the two subspecies after recent globalisation, illustrating higher migration rate from high-income to low-income countries, which facilitated gene flow, recombination, population admixture and selective sweep.

Genetic exchanges between *C. h. hominis* (clade 1) and *C. hominis aquapotentis* (clade 2, *gp60* subtype IbA10G2) are rare compared to those within *C. parvum parvum* (Corsi, et al. 2021), and their frequency might be more comparable to the rate of sequence exchange between *C. p. parvum* and *C. p. anthroponosum* (Nader, et al. 2019). However, whole genome analyses of more *C. hominis* isolates may detect other recombination events in addition to the four events detected in our study. In addition, further studies may be able to discover the unknown parental sequences associated with recombination event 1 in our study. Without this parental sequence, we were unable to reconstruct the evolution of this introgressed sequence on chromosome 1. Yet, this event may be a key player in the evolution of the lineage. We encourage future whole genome studies on *C. hominis*, believing this may shed further light on the incipient speciation of *C. hominis aquapotentis* (clade 2, *gp60* subtype IbA10G2).

Although our dataset comprises samples across five continents, we only studied *C. hominis* in 16 countries in total, which means that we could have missed local gene flow and patterns of population sub-structuring within continents. Although the marked biological differences between *C. parvum* and *C. hominis* have been well established (Abrahamsen, et al. 2004), recent population genomic research is demonstrating that also within these species, the population genetics and evolutionary genetics of their subspecies are remarkably distinct. Large datasets and comparative population genomic and phylogenomic analyses (across *Cryptosporidium* species) are warranted to examine the evolutionary genomics of these parasites in more detail.

Finally, we have discovered genomic islands of putative virulence genes (GIPVs) contributing to population diversification between *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) and *C. hominis hominis* (clade 1). These islands have experienced relatively elevated recombination rate which has enriched nucleotide variation under balancing selection and the acquisition of non-synonymous SNPs, consistent with virulence factors driving host-parasite interactions. Intriguingly, the most significant signals within these analyses are driven by *gp60*, a hypothetical protein (CHUDEA6_5270) and a cluster of mucin-like genes (CHUDEA2_430, CHUDEA2_440 and CHUDEA2_450) found on chromosome 2. These genes are consistently identified as being under selection in the evolution of the *C. hominis* subspecies here and in similar observations made of *C. hominis* in Africa (Tichkule, et al. 2021). Their orthologs are associated with recombination between human-specific *C. parvum anthroponosum* relative to the zoonotic *C. parvum parvum* and appear to have driven convergence of the former with *C. hominis* (Nader, et al. 2019).

CHUDEA2_430 (*muc5*) and hypothetical protein CHUDEA6_5270 are the most notable, displaying significant diversifying selection between the two subspecies. Broadly, mucins mediate cell-cell interactions (O'Connor, et al. 2009), and modulate infectivity of *Cryptosporidium* sporozoites and merozoites and oocyst production (Cevallos, et al. 2000; O'Connor, et al. 2009). In *C. parvum*, MUC5 is involved in host-cell invasion and an important determinant of host adaptation (O'Connor, et al. 2009) and highly expressed in the first 2 hours of infection in vitro (Lippuner, et al. 2018). MUC5 may also play a role in tethering the sporozoite to the oocyst wall (Chatterjee, et al. 2010). Our analyses suggest *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) and *C. h. hominis* (clade 1) *muc5* haplotypes diverged before or with the subspecies and subsequently diversified, which is consistent with prior observations implicating CHUDEA2_430 in the emergence of *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) (Bouzid, et al. 2013; Feng, et al. 2018). This appears to have resulted in the acquisition of novel glycosylation sites within *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) *muc5* haplotypes. We cannot determine the functional consequence of these sites but note that glycosylation sites often mediate the specificity of mucin interactions with host proteins in a variety of pathogens (Lin, et al. 2020). In contrast, CHUDEA6_5270 displays a clear signal for the recent introgression of a novel *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) haplotype into *C. h. hominis* (clade 1) after the divergence of these subspecies. This haplotype has notable, structurally relevant, mutations. Identifying the function of this gene, its potential role in infection and the relevance of the structural variation we have inferred here, should be considered a major research priority.

In conclusion, this work represents the first large scale population genomic study in any *Cryptosporidium* species, inferring the global population structure and evolutionary history of *C. hominis*. We propose recognition of two distinct subspecies, *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2), with distinct demographic histories that have diverged circa 500 years ago. Although the subspecies differ in their global distribution, their gene pools are not completely isolated, and rare genetic exchanges have occurred in the recent past. We contend that many of the genes, CHUDEA2_430 and CHUDEA6_5270 in particular, in these introgression regions are involved in infection, and that their evolution in humans may be driving greater human specificity, virulence and transmissibility. It appears *C. h. aquapotentis* (clade 2, *gp60* subtype IbA10G2) is playing a key role in this process, which is supported by previous observations based on multilocus typing (Li, et al. 2013). This illustrates how human-mediated gene flow is involved in parasite evolution and genomic architecture, and how it could affect virulence evolution. Also, it shows that the GIPVs that result from population admixture in an anthroponotic species are under selection and involved in evolutionary arms race.

Methods

Parasite isolates

The *C. hominis* isolates newly sequenced for this study (n = 34) were archived stool samples collected at the *Cryptosporidium* Reference Unit in the UK. The species was determined by species-specific real-time PCR targeting the A135 gene (Robinson, et al. 2020) and subtyped by PCR and sequencing of the *gp60* gene (Chalmers, et al. 2019). Supplementary Table 1 provides information about these isolates. Isolates were selected to mainly represent the dominant variant, IbA10G2, as defined by *gp60* sequencing.

Processing of faecal samples for whole genome sequencing

Stool samples were processed as previously described (Hadfield, et al. 2015). Briefly, saturated salt-flotation was used to obtain a partially purified suspension of oocysts starting from 1-2 ml of each faecal sample. Oocysts were further purified from the suspension by immunomagnetic separation (IMS), using the Isolate® IMS kit (TCS Biosciences, Botolph Claydon, UK). IMS-purified oocysts were treated with bleach, and washed three times with nuclease-free water by centrifugation at 1,100´ g for 5 min. The pellets were suspended in 200 µL of nuclease-free water for DNA extraction.

DNA preparation and whole genome sequencing

Genomic DNA was extracted from purified *Cryptosporidium* oocysts by first performing eight cycles of freezing in liquid nitrogen for 1 min and thawing at 95°C for 1 min, and then using the QIAamp DNA extraction kit (Qiagen, Manchester, UK) according to the manufacturer's instructions. The genomic DNA was eluted in 50 µL nuclease-free water, and the concentration measured using the Qubit dsDNA HS Assay Kit with the Qubit 1.0 fluorometer (Invitrogen, Paisley, UK), according to the manufacturer's instructions.

Whole genome amplification (WGA) was performed using the Repli-g Midi kit (Qiagen, Milan, Italy), according to the manufacturer's instructions. Briefly, 5 µL of genomic DNA (containing 1-10 ng of DNA) were mixed with 5 µL of denaturing solution and incubated at room temperature for 3 min. Next, 10 µL of stop solution were added to stabilise denatured DNA fragments. The reaction mixture was completed with 29 µL of buffer and 1 µL of phi29 polymerase, and allowed to proceed for 16 hours at 30°C. The reaction was stopped by heating at 63°C for 5 minutes. WGA products were visualised by electrophoresis on a 0.7% agarose gel, purified and quantified by Qubit as described above.

For Next Generation Sequencing (NGS) experiments, about 1 µg of purified WGA product was used to generate Illumina TruSeq 2x 150 bp paired-end libraries (average insert size: 500 bp), which were sequenced on an Illumina HiSeq 4000 platform (Illumina, SanDiego, CA). Library preparation and NGS experiments were performed by a commercial company (GATC, Germany).

Whole genome global dataset

To perform a global comparative genomics of *C. hominis*, we supplemented our newly sequenced genome dataset by downloading all available published *C. hominis* genome sequences on till date (25th July 2021), from the sequence read archive (SRA) of NCBI and from the EMBL's European Nucleotide Archive (ENA) (see Supplementary Table 1). Collectively, these data represented 114 genome sequences of locally acquired infections from 16 countries across five continents.

Data pre-processing and variant calling

Raw reads of the 114 *C. hominis* isolates were trimmed to remove adapter sequences and filtered for low-quality bases using Trimmomatic v.0.36 (Bolger, et al. 2014). The filtered reads were aligned to *C. hominis* UdeA01 reference genome (Heiges, et al. 2006; Isaza, et al. 2015) using the maximal exact matches (MEM) algorithm implemented in Burrows-Wheeler Alignment (BWA) tool v.0.7 (Li and Durbin 2009) with default settings. PCR duplicates were then marked using Picard MarkDuplicates (https://broadinstitute.github.io/picard/) followed by Genome Analysis Toolkit's (GATK) indel realignment and base quality score recalibration (BQSR) using default parameters (McKenna, et al. 2010). Sequence variants (SNPs) were called from the aligned reads of each isolate using the HaplotypeCaller method in the GATK v3.7.0 (McKenna, et al. 2010) as per GATK's best practices pipeline (Van der Auwera, et al. 2013). SNPs were removed if quality depth (QD) < 2.0, Fisher strand (FS) > 60.0, mapping quality (MQ) < 40.0, mapping quality rank sum test (MQRankSum) < -12.5, read position rank sum test (ReadPosRankSum) < -8.0, Strand odds ratio (SOR) > 4.0. All identified SNPs were combined in one file and each isolate genotyped using the GenotypeGVCFs tool (GATK v3.7.0) (McKenna, et al. 2010). To maximise the quality, SNPs were further filtered based on the following criteria and included in the downstream process: bi-allelic SNPs, quality score > 30, allele depth (AD) > 5, MAF > 0.05 and missing ratio < 0.5. Each of the 114 whole genome sequences assessed here had > 80% coverage of the *C. hominis* reference genome to at least the 5-fold depth. Each of the 114 whole genome sequences assessed here had at least ~80% coverage of the *C. hominis* reference genome to at least the 5-fold depth where 103/114 has > 80% genome coverage and at least 10X coverage. Mean coverage of all isolates is 158X (Quartile1 = 117X, Quartile3 = 229X)] (Supplementary Table 1).

Population genetic structure based on whole genome SNPs

The filtered bi-allelic SNPs were used for population structure, phylogenetic and clustering analyses. Multiplicity of infections in each sample were estimated using estMOI (Assefa, et al. 2014) and

MOIMIX (https://github.com/bahlolab/moimix). MOIMIX calculates *Fws* statistic (Manske, et al. 2012), a fixation index that is used to assess within-host genetic differentiation. An isolate with single infection is expected to have *Fws* 0.95 - 1.00. The R package SNPRelate v.1.18 (Zheng, et al. 2012) was used for principal-component analysis (PCA) analysis. seqVCF2GDS function in SNPRelate R package is used to first convert VCF file into genomic data structure (GDS) file format to store SNP genotypes in an array-oriented matrix format. A genetic covariance matrix is then calculated from genotypes using SNPRelate's function snpgdsPCA, along with the correlation coefficients between samples and genotypes for each SNP. A maximum likelihood phylogenetic tree was constructed by IQ-TREE (Nguyen, et al. 2014) with 1000 bootstraps and visualised in iTOL v3 (Letunic and Bork 2016); the sister species *C. parvum* was used as an outgroup. We also constructed a consensus of 10^7 trees using DensiTree 2 (Bouckaert and Heled 2014) in BEAST v2 (Bouckaert, et al. 2014). BEAST v2 (Bouckaert, et al. 2014) was also used to estimate the divergence time between the populations by using 95% highest posterior density (HPD); and SpeedDate (https://github.com/vanOosterhoutLab/SpeedDate.jl) to estimate the coalescence times between sequences by using 5-95% confidence interval (CI). We used mutation rate of 10^{-8} and a generation time of 48h/replication (Nader, et al. 2019) to date the coalescence times between sequences. A Neighbor-Net algorithm-based network was generated using SplitsTree5 (Huson and Bryant 2006). Genetic structure was analysed by STRUCTURE v2.3 software (Pritchard, et al. 2000) for population number (K) ranging 2 - 10 and plotted by using plotSTR R package (https://github.com/DrewWham/Genetic-Structure-Tools). The optimal population genetic cluster value K was estimated by using CLUMPAK (Kopelman, et al. 2015).

Population demographic history and divergence time estimation

We used Bayesian Markov Chain Monte Carlo (MCMC) method implemented in Beast v2 program (Bouckaert, et al. 2014) to estimate the effective population size (*Ne*) of the *C. hominis* population. The nucleotide substitution model of HKY was selected. A strict molecular clock model and a Bayesian skyline coalescent tree prior was used with 10^9 generations of MCMC chain and 10% burn-ins. Tracer v.1.7 (Rambaut, et al. 2018) was used to assess chain convergence and effective sample size [ESS] > 200 and to construct the demographic history over time; i.e. Bayesian Skyline Plot (BSP). SweeD (Pavlidis, et al. 2013) was used to detect windows of selective sweeps from genome-wide SNP dataset by using composite likelihood ratio (CLR) statistic that identifies signature of site frequency spectrum (SFS), with a grid size of 1000.

Demographic histories and migration rates between the *C. hominis* populations were estimated by using fastsimcoal2 (Excofffier, et al. 2021) by using mutation rate of 10^{-8} and a generation time of 48h/replication (Nader, et al. 2019). We first inferred best parameters and the likelihoods for each of the demographic models – no geneflow, ongoing geneflow, early geneflow, recent geneflow and different geneflow, since the time of divergence (~500 years) by running 100 independent iterations with 300,000 coalescent simulations and 60 optimisation cycles. Demographic model with the highest likelihood (log10) was then selected to run parameter estimation with block-bootstrapping of 100 replicates.

Linkage, recombination and gene flow analyses

We inferred the rate of decay of linkage disequilibrium (LD) by calculating the squared correlation of the coefficient (r^2) between SNPs within 50kb using VCFtools (Danecek, et al. 2011). LD blocks were also determined by calculating pairwise r^2 between SNPs within chromosomes of each population. Recombination events were identified using the Recombination Detection Program version 4 (RDP4) (Martin, et al. 2015) using the RDP (Martin and Rybicki 2000), GENECONV (Sawyer 1999), BootScan (Salminen, et al. 1995), MaxChi (Smith 1992) and Chimaera (Posada and Crandall 2001) methods. Events were considered significant if at least three methods predicted their occurrence at a probability values, $p \leq 10^{-5}$. Recombination events with undetermined parental sequences were excluded from further HybridCheck analyses. Statistically significant recombination events were visualised and analysed using HybridCheck (Ward and van Oosterhout 2016) to determine the sequence similarity between the isolates involved in the events. HybridCheck program was also used to calculate the D statistic and estimate the gene flow between the populations.

Population genetic and genomic analyses of coding region

Tajima's D, Nucleotide diversity (π), Dxy and Fst were calculated using the PopGenome R-package (Pfeifer, et al. 2014). Nonsynonymous (Ka) and synonymous (Ks) mutation rates were calculated by using Ka/Ks_Calculator (Zhang, et al. 2006). Protein localisation (extracellular) was predicted using WoLF PSORT (Horton, et al. 2007) and information regarding predicted protein targeting (signalling peptides) genes were obtained from CryptoDB (Heiges, et al. 2006). POPART program was used to generate haplotype networks (Leigh and Bryant 2015). AlphaFold was used to predict the protein structures (Jumper, et al. 2021). Glycosylation sites were predicted by using NetNGlyc 4.0 Server (Gupta and Brunak 2002). Intrinsically disordered region in proteins were predicted using IUPred2A (Mészáros, et al. 2018). All statistical tests and results were performed and plotted in R (version 3.6.1).

Data access

The raw data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers

PRJEB15112, PRJNA610731, PRJNA610732, PRJNA610735, PRJNA610737, PRJNA610738, PRJNA610739, PRJNA610740, PRJNA610741, PRJNA610742, PRJNA610743, PRJNA610744, PRJNA610745, PRJNA610746, PRJNA610747 and PRJNA610748. Reviewer link to deposited data for which the accessions are not yet public, are provided in the following table.

BioProject ID	Reviewer link
PRJNA610731	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610731?reviewer=i5832fbel5senhl9g2ju3rj1n1
PRJNA610732	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610732?reviewer=j2pdubv6o5ks7q8earef6kt38r
PRJNA610735	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610735?reviewer=g3v1kq0afu72ff6tcc1o61sj5n
PRJNA610737	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610737?reviewer=q4joev2kaeck045i21eq6i9s1n
PRJNA610738	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610738?reviewer=r4pf8pchijamuem60ncgk1ln01
PRJNA610739	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610739?reviewer=8rgo5b89r49hd0cl27gpmgg1o7
PRJNA610740	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610740?reviewer=rthk0pe2qsn5kc6rn2kdomgo17
PRJNA610741	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610741?reviewer=njvi5jgljq271l2i5ef6hlo0jl
PRJNA610742	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610742?reviewer=jdgqt0q8lkgtnt083sdnkrkj0c
PRJNA610743	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610743?reviewer=9qka47natbb858bm5fejj5q78q
PRJNA610744	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610744?reviewer=8vvg89inuefhack29eo5jscu9b
PRJNA610745	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610745?reviewer=lqam541l8qbhqephabfeq8mu72
PRJNA610746	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610746?reviewer=3hqil4t2u9phuucr8a5nlvnjk0
PRJNA610747	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610747?reviewer=a9bb6grp0g2vsmvljijsgig17t
PRJNA610748	https://dataview.ncbi.nlm.nih.gov/object/PRJNA610748?reviewer=1epesalvpd5qneoadvro9u2436

Acknowledgments

S.T. acknowledges the Australian Society for Parasitology (ASP) for a student conference travel grant, and JD Smyth Postgraduate Travel Award for a Network Researcher Exchange, Training and Travel Award; and Walter and Eliza Hall Research Institute (WEHI), Australia, for the top up scholarship. S.T. also acknowledges Namrata Srivastava (Monash University) for assisting with statistical data analysis.

Funding source

This work was supported by the Australian National Health and Medical Research Council (APP1194330) and Victorian State Government Operational Infrastructure Support and Australian Government National Health and Medical Research Council Independent Research Institute Infrastructure Support Scheme to A.R.J. This work was supported by the Interreg 2 Seas programme 2014–2020 co-funded by the European Regional Development Fund under subsidy contract No.

2S05–043 H4DC to K.M.T. This work was supported by the European Union's Horizon 2020 research and innovation programme, project “Collaborative management platform for detection and analyses of (re-) emerging and foodborne outbreaks in Europe” (COMPARE, www.compare-europe.eu), grant agreement No. 643476 to S.M.C. This work was supported by Walter and Eliza Hall International PhD Scholarship and Melbourne Research Scholarship (MRS) to S.T. This work was supported by the Earth and Life Systems Alliance (ELSA) of the Norwich Research Park (NRP) to C.V.O.

Competing interests

The authors declare that there are no conflicts of interest.

Author contributions

A.R.J., S.M.C., C.V.O. and S.T. conceived the study. A.R.J., S.M.C., C.V.O. and S.T. designed the analyses. S.M.C., R.M.C., G.R., D.E., and K.M.T. were involved in acquisition of data. S.T. performed the bioinformatics associated evolutionary genetic and genomic analyses. A.R.J., S.M.C., C.V.O. and S.T. wrote the manuscript. All authors read and approved the submission of the manuscript for the publication.

Figure Legends

Fig 1. Global population structure of *C. hominis* isolates illustrating their sub-structing and diversification. **a**. PCA of isolates based on the filtered set of whole genome SNPs, highlighting three clusters of isolates which are predominately based on continents of origin. Isolates were color coded with their continent of origin. Isolates associated with *gp60* subtype IbA10G2 were represented with solid circles while non-IbA10G2 with solid triangles. **b.** Structure plot illustrating population genetic ancestry and the admixed nature of the *C. hominis* isolates. The plot was obtained for an optimum value of K=4. The black arrow (bottom) indicates the highly admixed isolate (UK_UKH4), which includes all four ancestries. **c.** Maximum likelihood based phylogenetic tree. **d.** Splitstree and **e.** Densitree are also demonstrating two major clades. *C. h. hominis* (clade 1) includes isolates associated with other *gp60* subtypes while *Cryptosporidium h. aquapotentis* (clade 2) includes isolates associated with *gp60* subtype IbA10G2.

Fig 2. Demographic histories and population size and secondary contact between *C. h. hominis* (clade 1) and *C. h. aquapotentis (*clade 2). **a**. Bayesian Skyline plots (BSP) depicting change in *Ne (*effective population size) through time, for both the clades. The central dark line and the upper and lower dashed lines on Y-axis are mean estimates and 95% HPD intervals of *Ne*, respectively. X-axis is time in years, running backwards. **b**. Boxplot showing significant difference (two-sided t-test) in Tajima’s

D values between *C. h. hominis (*clade 1) and *C. h. aquapotentis (*clade 2). **c**. Higher likelihood (log10) for "recent geneflow" model (in red). Comparing likelihood distributions of geneflow models and observed significant difference (one-way ANOVA test, F = 2629761, df = 4, p-value <2e-16). Further, Post-hoc Tukey-HSD test revealed difference in likelihood between all the models (p-value < 1e-16). **d**. Graphical representation of demographic history of *C. hominis*, illustrating recent secondary contact and migration rates between the two clades (mean ± SE).

Fig 3. Analyses of recombination and gene flow between *C. h. hominis* (clade 1) and *C. h. aquapotentis (*clade 2). **a**. Linkage disequilibrium (LD) decay plot showing rapid decay of linkage between SNPs in *C. h. hominis* (clade 1) compared to *C. h. aquapotentis (*clade 2). **b**. Graphical representation of recombinant breakpoint positions detected by RDP4 program between *C. h. hominis* (clade 1) and *C. h. aquapotentis (*clade 2). **c-d**. HybridCheck plots representing genomic signature of introgression in chromosome 2 and 6, respectively. Analysis for chromosome 1 was excluded due to unknown parental sequences. The plots were generated for random set of triplets that includes recombinant (hybrid), minor (donor) and major (recipient) parental sequence, as detected by RDP4 program. Introgressed blocks (recombinant breakpoints) were illustrated with dashed boxes, showing high similarity between the recombinant (*C. h. hominis* hybrid isolates) and minor parent (*C. h. aquapotentis* isolates). The top panel illustrates the visualisation of sequence similarity between sequences within the triplet, using RBG colour triangle. The two sequences are coloured same (yellow, purple or turquoise) if they share polymorphism. **e.** Gene flow analyses with ABBA-BABA test, representing D statistics for the random sets of triplets (as used in c-d) along with *C. parvum* as an outgroup. D statistic values close to -1 at all three recombinant events, suggesting geneflow between H1 and H3. **f-g**. Pairwise LD of SNPs in chromosomes 2 and 6 of *C. h. hominis* showing red-blocks of high linkage between SNPs in introgressed events 2-4.

Fig 4. Population genetic analyses of genomic islands of putative virulence genes (GIPVs). **a**. Population genetic and divergence analyses of introgressed regions. X-axis represents genomic positions of eight chromosomes highlighted with different colours. Population divergence (Dxy) between *C. h. hominis (*clade 1) and *C. h. aquapotentis (*clade 2) for each gene were plotted on Y-axis (top panel). Nucleotide diversity (π) for *C. h. hominis* (middle panel) and *C. h. aquapotentis* (bottom panel) for each gene, were also plotted on Y-axis, respectively. The breakpoints of four recombination events (event 1-4) were indicated by grey vertical boxes. Event 1 was un-detected in *C. h. aquapotentis*. **b**. Correlation between π and Dxy were plotted to identify polymorphic and potential virulence genes.

Fig 5. Illustrating diversifying selection between *C. hominis* subspecies and host adaptation at CHUDEA6_5270 (hypothetical gene). **a.** Haplotype network analyses illustrating haplotype diversification between *C. h. hominis (*clade 1) and *C. h. aquapotentis (*clade 2). **b**. Pairwise nucleotide divergence shows bimodal distribution, which, theoretically, can be explained both by balancing selection (Lighten, et al. 2017), as well as by genetic introgression. **c.** Comparison of predicted models of protein structure of CHUDEA6_5270 gene between *Cryptosporidium* species and subtypes demonstrates variation towards C.-terminal region. **d**. Introgressed-isolates driving balancing selection at gene CHUDEA6_5270 in *C. h. hominis*. Red line represents balancing selection (positive Tajima's D) in *C. h. hominis* that also includes introgressed-isolates. Blue line represents purifying selection (negative Tajima's D) in *C. h. hominis* after excluding introgressed-isolates.

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, et al. 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science 304:441-445.

Assefa SA, Preston MD, Campino S, Ocholla H, Sutherland CJ, Clark TG. 2014. estMOI: estimating multiplicity of infection using parasite deep sequencing data. Bioinformatics 30:1292-1294.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLOS Computational Biology 10:e1003537.

Bouckaert RR, Heled J. 2014. DensiTree 2: Seeing Trees Through the Forest. bioRxiv:012401.
Bouzid M, Hunter PR, Chalmers RM, Tyler KM. 2013. *Cryptosporidium* pathogenicity and virulence. Clin Microbiol Rev 26:115-134.

Cacciò SM, Chalmers RM. 2016. Human cryptosporidiosis in Europe. Clin Microbiol Infect 22:471-480.

Cama VA, Bern C, Roberts J, Cabrera L, Sterling CR, Ortega Y, Gilman RH, Xiao L. 2008. *Cryptosporidium* species and subtypes and clinical manifestations in children, Peru. Emerging infectious diseases 14:1567.

Carmicheal J, Atri P, Sharma S, Kumar S, Chirravuri Venkata R, Kulkarni P, Salgia R, Ghersi D, Kaur S, Batra SK. 2020. Presence and structure-activity relationship of intrinsically disordered regions across mucins. The FASEB Journal 34:1939-1957.

Cevallos AM, Bhat N, Verdon R, Hamer DH, Stein B, Tzipori S, Pereira ME, Keusch GT, Ward HD. 2000. Mediation of *Cryptosporidium parvum* infection in vitro by mucin-like glycoproteins defined by a neutralizing monoclonal antibody. Infect Immun 68:5167-5175.

Chalmers RM, Robinson G, Elwin K, Elson R. 2019. Analysis of the *Cryptosporidium* spp. and *gp60* subtypes linked to human outbreaks of cryptosporidiosis in England and Wales, 2009 to 2017. Parasites & Vectors 12:95.

Chalmers RM, Robinson G, Elwin K, Hadfield SJ, Thomas E, Watkins J, Casemore D, Kay D. 2010. Detection of *Cryptosporidium* species and sources of contamination with *Cryptosporidium hominis* during a waterborne outbreak in north west Wales. Journal of water and health 8:311-325.

Chatterjee A, Banerjee S, Steffen M, O'Connor RM, Ward HD, Robbins PW, Samuelson J. 2010. Evidence for mucin-like glycoproteins that tether sporozoites of *Cryptosporidium parvum* to the inner surface of the oocyst wall. Eukaryotic cell 9:84-96.

Corsi GI, Tichkule S, Sannella AR, Vatta P, Asnicar F, Segata N, Jex AR, Oosterhout Cv, Cacciò SM. 2021. Evolutionary epidemiology of a zoonosis. bioRxiv:2021.2010.2015.464618.

Craun GF, Hubbs SA, Frost F, Calderon RL, Via SH. 1998. Waterborne outbreaks of cryptosporidiosis. Journal AWWA 90:81-91.

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.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22:1185-1192.

Efstratiou A, Ongerth JE, Karanis P. 2017. Waterborne transmission of protozoan parasites: review of worldwide outbreaks-an update 2011–2016. Water research 114:14-22.

Excofffier L, Marchi N, Marques DA, Matthey-Doret R, Gouy A, Sousa VC. 2021. fastsimcoal2: demographic inference under complex evolutionary scenarios. Bioinformatics.

Feng Y, Ryan UM, Xiao L. 2018. Genetic Diversity and Population Structure of *Cryptosporidium*. Trends Parasitol.

Feng Y, Tiao N, Li N, Hlavsa M, Xiao L. 2014. Multilocus sequence typing of an emerging *Cryptosporidium hominis* subtype in the United States. J Clin Microbiol 52:524-530.

Guo Y, Tang K, Rowe LA, Li N, Roellig DM, Knipe K, Frace M, Yang C, Feng Y, Xiao L. 2015. Comparative genomic analysis reveals occurrence of genetic recombination in virulent *Cryptosporidium hominis* subtypes and telomeric gene duplications in *Cryptosporidium parvum*. BMC Genomics 16:320.

Gupta R, Brunak S. 2002. Prediction of glycosylation across the human proteome and the correlation to protein function. Pac Symp Biocomput:310-322.

Hadfield SJ, Pachebat JA, Swain MT, Robinson G, Cameron SJ, 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.

Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, Kaluskar N, Rhodes P, Wang S, He CZ, Su Y, et al. 2006. CryptoDB: a *Cryptosporidium* bioinformatics resource update. Nucleic Acids Res 34:D419-422.

Herges GR, Widmer G, Clark ME, Khan E, Giddings CW, Brewer M, McEvoy JM. 2012. Evidence that *Cryptosporidium parvum* populations are panmictic and unstructured in the Upper Midwest of the United States. Appl Environ Microbiol 78:8096-8101.

Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. 2007. WoLF PSORT: protein localization predictor. Nucleic Acids Res 35:W585-W587.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254-267.

Isaza JP, Galván AL, Polanco V, Huang B, Matveyev AV, Serrano MG, Manque P, Buck GA, Alzate JF. 2015. Revisiting the reference genomes of human pathogenic *Cryptosporidium* species: reannotation of *C. parvum* Iowa and a new *C. hominis* reference. Sci Rep 5:16324.

Jex AR, Gasser RB. 2010. Genetic richness and diversity in *Cryptosporidium hominis* and *C. parvum* reveals major knowledge gaps and a need for the application of "next generation" technologies--research review. Biotechnol Adv 28:17-26.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature.

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.

King P, Robinson G, Elwin K, Tyler KM, Hunter PR, Chalmers RM. 2017. Prevalence and epidemiology of human *Cryptosporidium parvum* IIc infections in England and Wales. The Lancet 389:S56.

King P, Tyler KM, Hunter PR. 2019. Anthroponotic transmission of *Cryptosporidium parvum* predominates in countries with poorer sanitation: a systematic review and meta-analysis. Parasites & Vectors 12:16-16.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Molecular ecology resources 15:1179-1191.

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.

Leigh JW, Bryant D. 2015. popart: full-feature software for haplotype network construction. Methods in Ecology and Evolution 6:1110-1116.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242-245.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Li N, Xiao L, Cama VA, Ortega Y, Gilman RH, Guo M, Feng Y. 2013. Genetic recombination and *Cryptosporidium hominis* virulent subtype IbA10G2. Emerg Infect Dis 19:1573-1582.

Lighten J, Papadopulos AST, Mohammed RS, Ward BJ, G. Paterson I, Baillie L, Bradbury IR, Hendry AP, Bentzen P, van Oosterhout C. 2017. Evolutionary genetics of immunological supertypes reveals two faces of the Red Queen. Nature Communications 8:1294.

Lin B, Qing X, Liao J, Zhuo K. 2020. Role of Protein Glycosylation in Host-Pathogen Interaction. Cells 9:1022.

Lippuner C, Ramakrishnan C, Basso WU, Schmid MW, Okoniewski M, Smith NC, Hässig M, Deplazes P, Hehl AB. 2018. RNA-Seq analysis during the life cycle of *Cryptosporidium parvum* reveals significant differential gene expression between proliferating stages in the intestine and infectious sporozoites. Int J Parasitol 48:413-422.

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.

Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16:562-563.

Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evolution 1.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20:1297-1303.

McKerr C, Chalmers RM, Elwin K, Jones H, Vivancos R, O'Brien SJ, Christley RM. 2021. Cross-Sectional Household Transmission Study of *Cryptosporidium* Shows that *C. hominis* Infections are A Key Risk Factor for Spread.

Mészáros B, Erdős G, Dosztányi Z. 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res 46:W329-W337.

Morrison LJ, Mallon ME, Smith HV, MacLeod A, Xiao L, Tait A. 2008. The population structure of the *Cryptosporidium parvum* population in Scotland: a complex picture. Infect Genet Evol 8:121-129.

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*. Nature Microbiology.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 32:268-274.

Nichols GL, Chalmers RM, Hadfield SJ. 2014. Molecular epidemiology of human cryptosporidiosis. In. *Cryptosporidium*: parasite and disease: Springer. p. 81-147.

O'Connor RM, Burns PB, Ha-Ngoc T, Scarpato K, Khan W, Kang G, Ward H. 2009. Polymorphic mucin antigens CpMuc4 and CpMuc5 are integral to *Cryptosporidium parvum* infection in vitro. Eukaryotic cell 8:461-469.

Pavlidis P, Živkovic D, Stamatakis A, Alachiotis N. 2013. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol Biol Evol 30:2224-2234.

Perez-Vilar J, Hill RL. 1999. The Structure and Assembly of Secreted Mucins* 210. Journal of Biological Chemistry 274:31751-31754.

Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol 31:1929-1936.

Pond SL, Frost SD. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531-2533.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proceedings of the National Academy of Sciences 98:13757.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Putignani L, Menichella D. 2010. Global distribution, public health and clinical impact of the protozoan pathogen *Cryptosporidium*. Interdiscip Perspect Infect Dis 2010.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol 67:901-904.

Razakandrainibe R, Diawara EHI, Costa D, Le Goff L, Lemeteil D, Ballet JJ, Gargala G, Favennec L. 2018. Common occurrence of *Cryptosporidium hominis* in asymptomatic and symptomatic calves in France. PLoS Negl Trop Dis 12:e0006355.

Robinson G, Elwin K, Chalmers RM. 2020. *Cryptosporidium* Diagnostic Assays: Molecular Detection. Methods Mol Biol 2052:11-22.

Ryan U, Zahedi A, Paparini A. 2016. *Cryptosporidium* in humans and animals—a one health approach to prophylaxis. Parasite Immunol 38:535-547.

Salminen MO, CARR JK, BURKE DS, McCUTCHAN FE. 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS research and human retroviruses 11:1423-1425.

Santin M. 2020. *Cryptosporidium* and *Giardia* in Ruminants. Vet Clin North Am Food Anim Pract 36:223-238.

Sawyer S. 1999. GENECONV: a computer package for the statistical detection of gene conversion. http://www.math.wustl.edu/~ sawyer.

Segura R, Prim N, Montemayor M, Valls ME, Muñoz C. 2015. Predominant virulent IbA10G2 subtype of *Cryptosporidium hominis* in human isolates in Barcelona: a five-year study. PLoS One 10:e0121753.

Smith JM. 1992. Analyzing the mosaic structure of genes. J Mol Evol 34:126-129.

Tichkule S, Jex AR, van Oosterhout C, Sannella AR, Krumkamp R, Aldrich C, Maiga-Ascofare O, Dekker D, Lamshöft M, Mbwana J, et al. 2021. Comparative genomics revealed adaptive admixture in *Cryptosporidium hominis* in Africa. Microb Genom 7.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11.10.11-33.

Van Oosterhout C. 2021. Mitigating the threat of emerging infectious diseases; a coevolutionary perspective. Virulence 12:1288-1295.

Wang R, Zhang L, Axen 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.

Ward BJ, van Oosterhout C. 2016. HYBRIDCHECK: software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. Mol Ecol Resour 16:534-539.

Widerström M, Schönning C, Lilja M, Lebbad M, Ljung T, Allestam G, Ferm M, Björkholm B, Hansen A, Hiltula J. 2014. Large outbreak of *Cryptosporidium hominis* infection transmitted through the public water supply, Sweden. Emerging infectious diseases 20:581.

Yang X, Guo Y, Xiao L, Feng Y. 2021. Molecular epidemiology of human cryptosporidiosis in low- and middle-income countries. Clinical Microbiology Reviews 34:e00087-00019.

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 4:259-263.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28:3326-3328.

Zhou L, Singh A, Jiang J, Xiao L. 2003. Molecular surveillance of *Cryptosporidium* spp. in raw wastewater in Milwaukee: implications for understanding outbreak occurrence and transmission dynamics. J Clin Microbiol 41:5254-5257.

Figures

Figure 1.

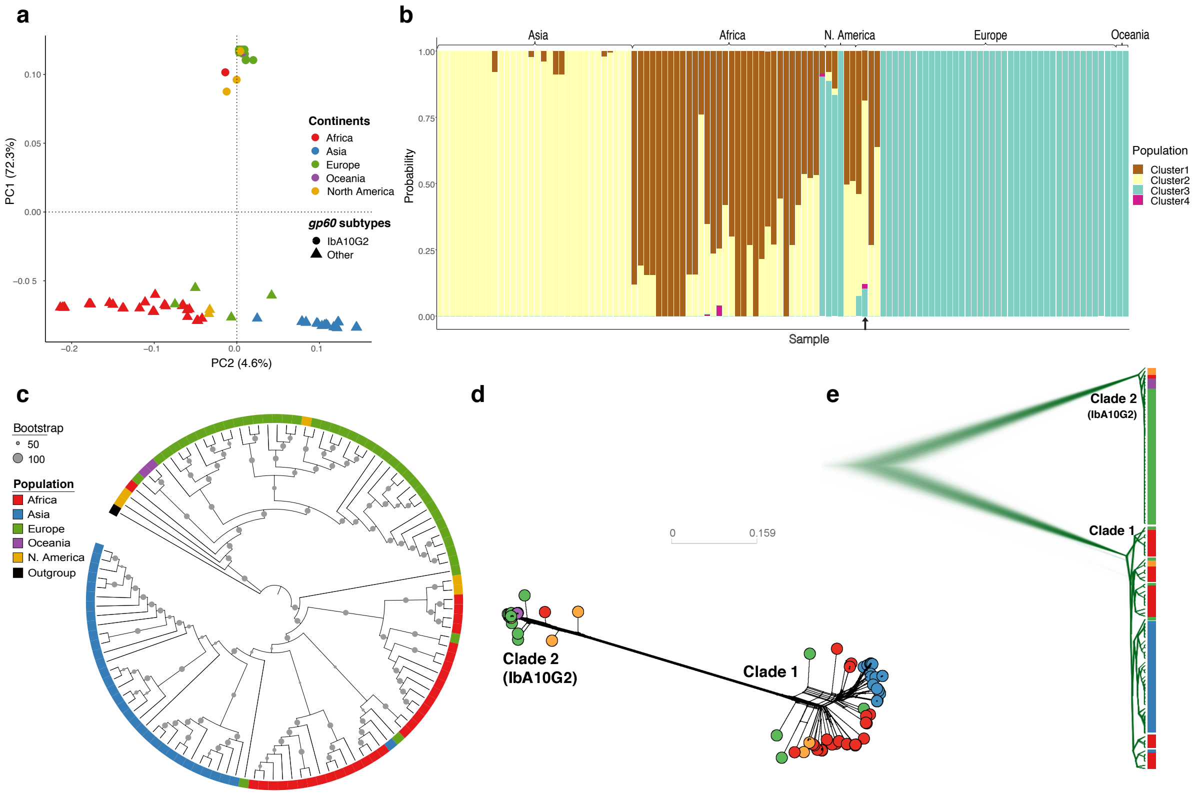


Figure 2.

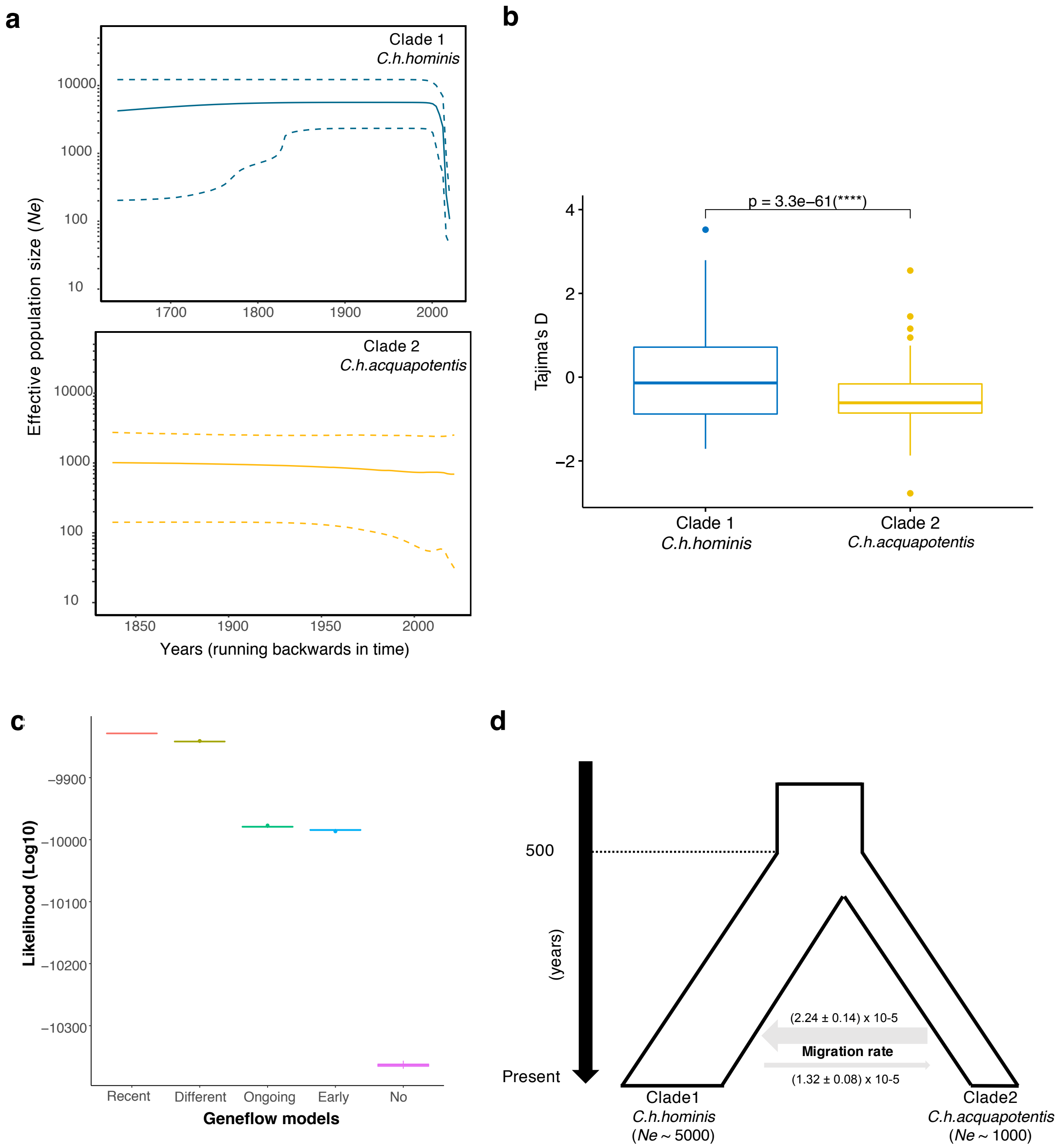
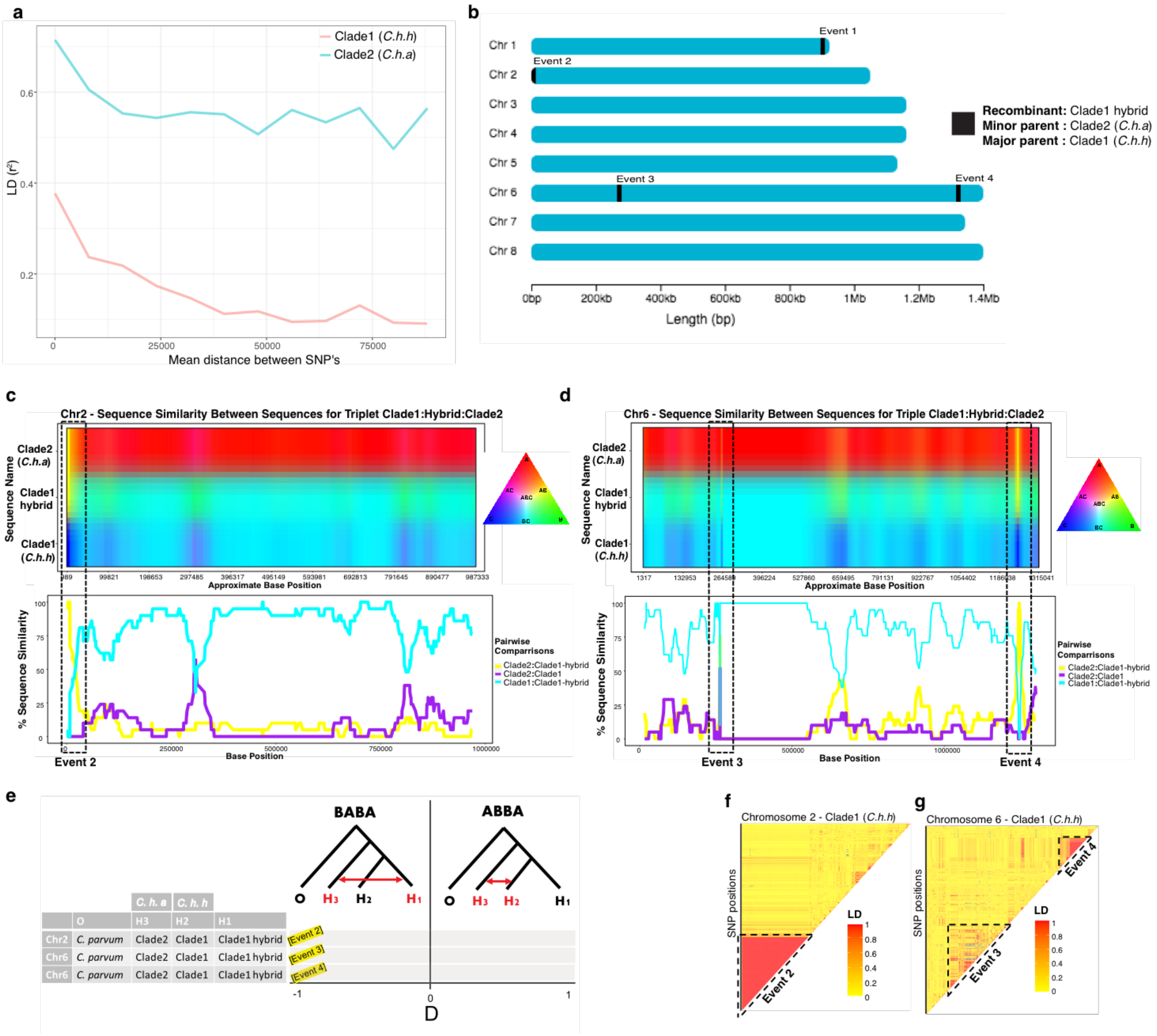

a
Clade 1
C.h.hominis
10000
1000
100
10
1700
1800
1900
2000
Clade 2
C.h.acquapotentis
10000
1000
100
10
1850
1900
1950
2000
Effective population size (Ne)
Years (running backwards in time)
b
p = 3.3e−61(****)
4
2
0
−2
Tajima's D
Clade 1
C.h.hominis
Clade 2
C.h.acquapotentis
c
−9900
−10000
−10100
−10200
−10300
Likelihood (Log10)
Recent
Different
Ongoing
Early
No
Geneflow models
d
500
(years)
Present
(2.24 ± 0.14) x 10-5
Migration rate
(1.32 ± 0.08) x 10-5
Clade1
C.h.hominis
(Ne ~ 5000)
Clade2
C.h.acquapotentis
(Ne ~ 1000)

Figure 3.



		C. h. a	C. h. h		
	O	H3	H2	H1	
Chr2	*C. parvum*	Clade2	Clade1	Clade1 hybrid	[Event 2]
Chr6	*C. parvum*	Clade2	Clade1	Clade1 hybrid	[Event 3]
Chr6	*C. parvum*	Clade2	Clade1	Clade1 hybrid	[Event 4]

Figure 4.

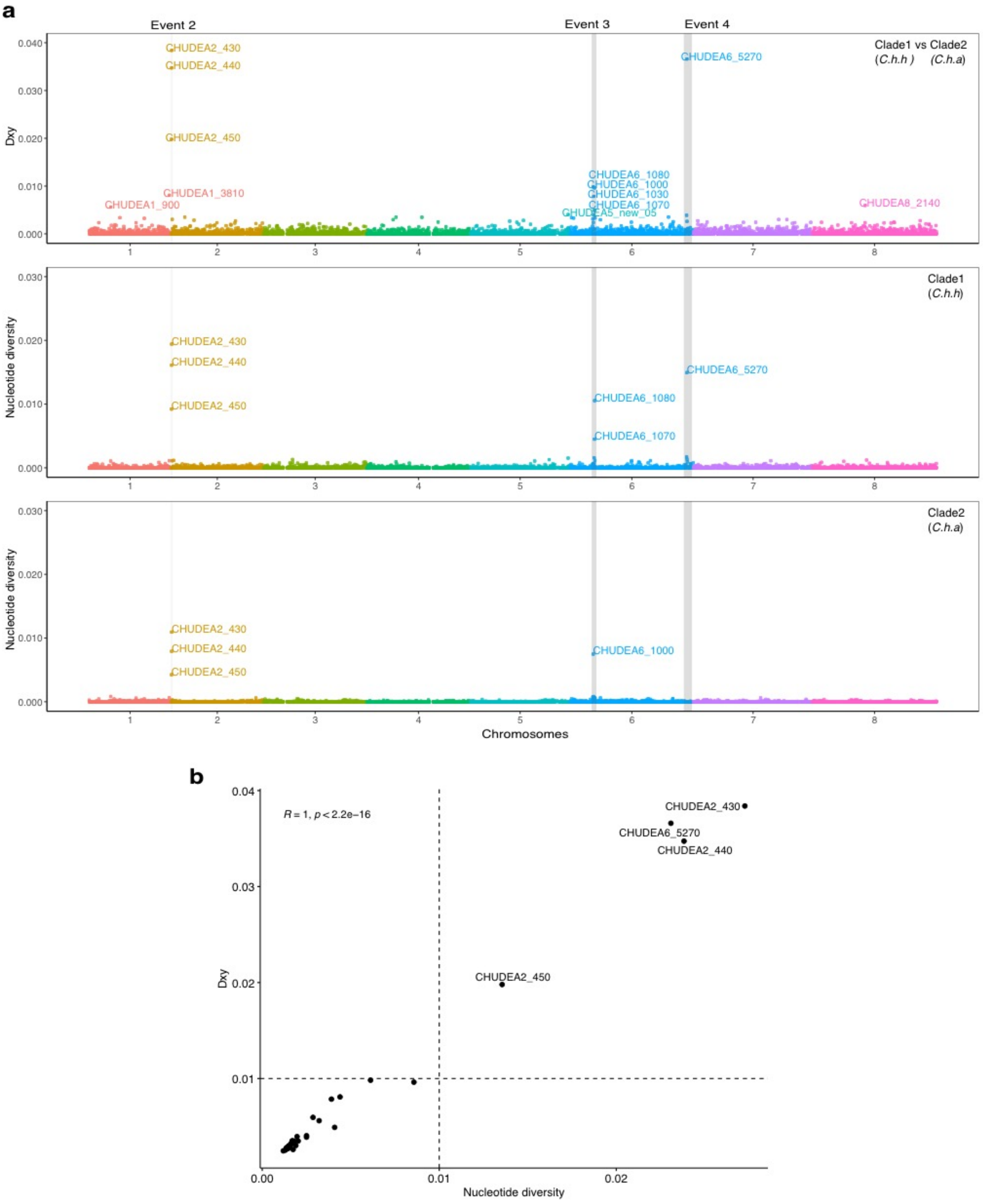
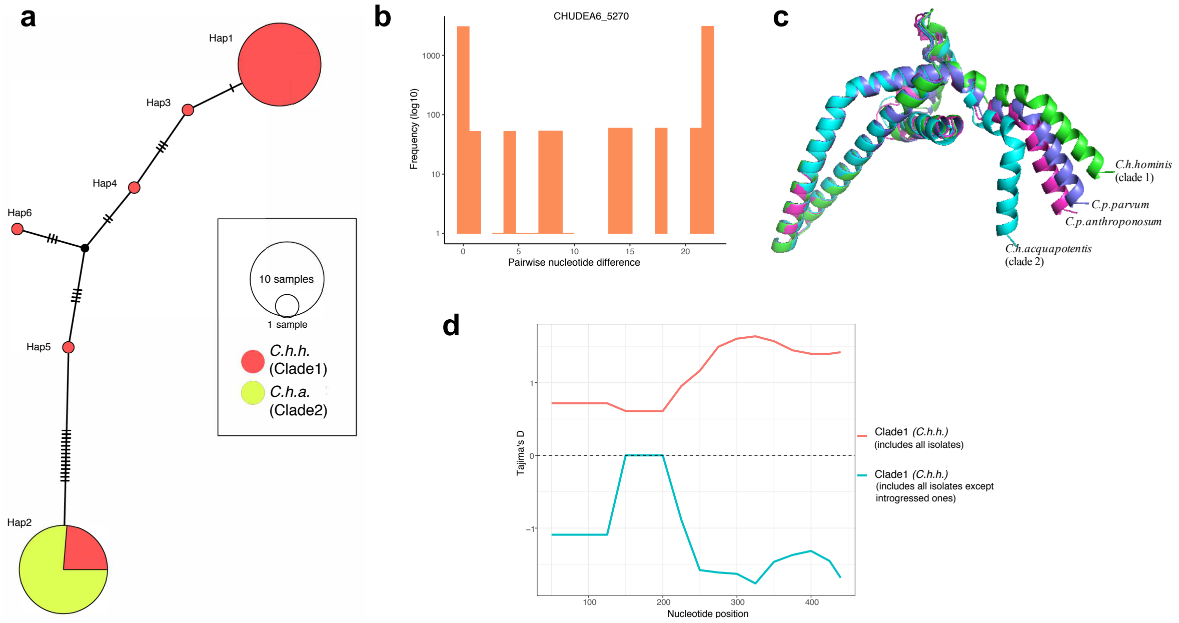

a
Event 2
Event 3
Event 4
Clade1 vs Clade2
(*C.h.h*) (*C.h.a*)
Dxy
0.040
0.030
0.020
0.010
0.000
CHUDEA2_430
CHUDEA2_440
CHUDEA2_450
CHUDEA1_3810
CHUDEA1_900
CHUDEA6_5270
CHUDEA6_1080
CHUDEA6_1000
CHUDEA6_1030
CHUDEA6_1070
CHUDEA5_new_05
CHUDEA8_2140
1
2
3
4
5
6
7
8
Clade1
(*C.h.h*)
Nucleotide diversity
0.030
0.020
0.010
0.000
CHUDEA2_430
CHUDEA2_440
CHUDEA2_450
CHUDEA6_5270
CHUDEA6_1080
CHUDEA6_1070
Clade2
(*C.h.a*)
Nucleotide diversity
0.030
0.020
0.010
0.000
CHUDEA2_430
CHUDEA2_440
CHUDEA2_450
CHUDEA6_1000
Chromosomes
b
R = 1, p < 2.2e−16
Dxy
0.04
0.03
0.02
0.01
CHUDEA2_430
CHUDEA6_5270
CHUDEA2_440
CHUDEA2_450
0.00
0.01
0.02
Nucleotide diversity

Figure 5.


a
Hap1
Hap3
Hap4
Hap6
Hap5
Hap2
10 samples
1 sample
C.h.h.
(Clade1)
C.h.a.
(Clade2)
b
CHUDEA6_5270
Frequency (log10)
1000
100
10
1
0
5
10
15
20
Pairwise nucleotide difference
c
C.h.hominis
(clade 1)
C.p.parvum
C.p.anthroponosum
C.h.acquapotentis
(clade 2)
d
Tajima's D
1
0
−1
100
200
300
400
Nucleotide position
Clade1 (C.h.h.)
(includes all isolates)
Clade1 (C.h.h.)
(includes all isolates except
introgressed ones)