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

E-mail addresses:

Swapnil Tichkule: tichkule.s@wehi.edu.au

Simone M. Cacciò: simone.caccio@iss.it

Guy Robinson: Guy.Robinson@wales.nhs.uk

Rachel M. Chalmers: Rachel.Chalmers@wales.nhs.uk

Ivo Mueller: mueller@wehi.edu.au

Samantha J. Emery-Corbin: emery.s@wehi.edu.au

Daniel Eibach: eibach@bnitm.de

Kevin Tyler: k.tyler@uea.ac.uk

Cock van Oosterhout: c.van-oosterhout@uea.ac.uk

Aaron Jex: jex.a@wehi.edu.au

*** Corresponding authors:**

E-mail: tichkule.s@wehi.edu.au (ST)

E-mail: c.van-oosterhout@uea.ac.uk (CVO)

E-mail: jex.a@wehi.edu.au (ARJ)

E-mail: simone.caccio@iss.it (SMC)

Keywords: *Cryptosporidium hominis*; speciation; whole genome sequencing; comparative genomics; population structure; population genetics; recombination; gene flow; secondary contact; selective sweep; evolution.

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

70 subspecies. This resulted in high diversity and divergence in genomic islands of putative
71 virulence genes (GIPVs), including *muc5* (CHUDEA2_430) and a hypothetical protein
72 (CHUDEA6_5270). This diversity is maintained by balancing selection, suggesting a
73 coevolutionary arms race with the host. Lastly, we find that recent gene flow from *C. h.*
74 *aquapotentis* to *C. h. hominis*, likely associated with increased human migration, may be
75 driving evolution of more virulent *C. hominis* variants.
76
77 **Main**
78 Cryptosporidiosis is a leading cause of diarrhoea in children under five globally[1,2], resulting
79 in an estimated 48,000 deaths annually. Among parasitic diseases, it is second only to malaria
80 in global health burden, with an overall impact of ~12.8 million Disability Adjusted Life-
81 Years (DALYs)[1]. Human cryptosporidiosis is primarily caused by *Cryptosporidium parvum*,
82 a zoonoses common in young ruminants[3,4], and *C. hominis*, which is anthroponotic and the
83 more prevalent species globally[5]. The disease burden is overwhelmingly skewed to low and
84 middle-income countries (LMICs)[6], particularly in sub-Saharan Africa[1], where *C. hominis*
85 predominates. However, *Cryptosporidium* remains a significant public health problem in
86 wealthy countries through large water or foodborne outbreaks and direct transmission in day-
87 care facilities, hospitals and other institutions[7,8]. Presently, there are no vaccines or effective
88 drugs to treat the infection. Hence, control depends on prevention of infection, which is
89 driven by a strong understanding of the parasite's epidemiology.
90 The global epidemiology of cryptosporidiosis varies by geographic region,
91 socioeconomic status and a range of risk factors[6,9]. Understanding of this epidemiology is
92 underpinned by molecular typing, based mainly on the highly polymorphic *gp60* gene[10]. This
93 work has identified numerous genetic variants within each species and indicated a complex
94 population genetic structuring. In *C. parvum*, population structure varies globally from clonal

95 to epidemic to panmictic, likely due to varying ecologic factors[11-13]. Genetic variants found
96 exclusively in humans point to an anthroponotic *C. parvum* lineage (IIc)[14,15], with a recent
97 genomic study recognising two subspecies, the zoonotic *C. parvum parvum* and the
98 anthroponotic *C. parvum anthroponosum*[16]. Interestingly, this study found that both
99 subspecies occasionally still hybridise and exchange genetic variation. These exchanges
100 overlapped with similar genomic regions undergoing genetic introgression between *C.*
101 *hominis* and *C. parvum anthroponosum*, indicating candidate sites underpinning adaptation to
102 human-specific infection[16]. *Cryptosporidium hominis* appears to have a largely clonal
103 population structure, dominated by specific variants in different regions[17]. The IbA10G2
104 *gp60* subtype, although found in LMICs[18], is the dominant variant in high-income countries
105 and accounts for up to ~45% of all *gp60*-typed human infections[18]. This subtype is linked to
106 most major waterborne outbreaks in wealthy countries for which genetic typing is available[19-
107 23]. The IbA10G2 subtype also appears more readily capable of direct human-to-human
108 transmission[24] and may cause more severe disease[25].

109 Studies of *C. hominis* molecular epidemiology pose several essential questions that
110 have major implications for global control of cryptosporidiosis. Specifically: (1) what is the
111 IbA10G2 subtype, and is this *gp60*-defined subtype reflective of a phylogenomically
112 divergent *C. hominis* lineage that predominates in wealthy countries; (2) if so, does this
113 lineage undertake reduced levels of genetic recombination with global *C. hominis*
114 populations; (3) and can signatures within the genomes IbA10G2 typed isolates identify its
115 taxonomic status, reveal the factors underpinning its increased virulence and its dominance in
116 wealthy countries, and identify its influence on global parasite population structure? To
117 address these questions, we performed a global study of 114 *C. hominis* genomes, comparing
118 the IbA10G2 subtype to published genomes representing locally acquired infections from 16
119 countries across five continents. The insights gained from these analyses are particularly

120 relevant for public health; the IbA10G2 subtype has already been identified as an emerging
121 host-adapted, virulent, and hyper-transmissible population, making it a threat to human health
122 in high-income countries[26-28].
123
124 **Results**
125 **Genomic evidence of population sub-structuring at the continental level**
126 All 114 *C. hominis* isolates included in this study were confirmed as single variant infections
127 (estimated MOI=1). We identified 5,618 biallelic SNPs among these samples and used these
128 to explore the population structure of *C. hominis*. Our analyses identified two major clusters
129 (Fig. 1a), separating European, North American and Oceanian samples from Asian and
130 African samples. We also saw minor clustering separating African from Asian samples.
131 STRUCTURE analysis provided further support for these findings (Fig. 1b). Overall, ~94%
132 of isolates clustered geographically. Exceptions included a small number of infections (e.g.,
133 UKH30 (from UK) and Sweden_SRR3098109 (from Sweden) acquired during international
134 travel. STRUCTURE analysis also identified a fourth cluster, including three African and one
135 European isolate, with unique population ancestry (cluster 4 in Fig. 1b).
136
137 **Genomic evidence of population diversification**
138 Maximum Likelihood (ML) analysis (Fig. 1c) identified two major clades, one corresponding
139 to Asia and Africa (clade 1) and the other to Europe, North America and Oceania (clade 2).
140 All isolates within clade 2 had *gp60* subtype IbA10G2, and no IbA10G2 isolates clustered
141 with clade 1. The two clades were estimated to have diverged 488 (84 – 2,199; 95% HPD)
142 years ago. This divergence is supported by a network (Fig. 1d) and SplitTree and DensiTree
143 (Fig. 1e) analyses. Considering the stability of these two genomically distinct lineages and

144 evidence below of their reproductive isolation, we propose their recognition as separate
145 subspecies. We propose clade 1, which comprises infections observed in low- to middle-
146 income countries, mostly from Africa and Asia, be recognised as *C. hominis hominis*,
147 referring to the fact that this subspecies represents the majority population. We propose clade
148 2 (IbA10G2 subtype) includes isolates from high-income countries, namely Europe, North
149 America and Oceania, be named *C. hominis aquapotentis* (strong water), as it predominates
150 in countries with longstanding high sanitation and water quality indices.

151

152 **Demographic histories**

153 To understand the demographic histories and estimate the change in the effective population
154 size (*Ne*) through time, we constructed a Bayesian Skyline Plot (BSP)[29] for *C. h. hominis* and
155 *C. h. aquapotentis*. Parasite isolates from low-income countries (*C. h. hominis*) experienced a
156 marked population contraction recently from *Ne*=~5000 to *Ne*=~100 (Fig. 2a), which is
157 supported by a higher proportion of positive Tajima's D values (Fig. 2b). In contrast, the
158 isolates from high-income countries (*C. h. aquapotentis*, *gp60* subtype IbA10G2) had a stable
159 effective population size (*Ne*=~1000) and a higher proportion of negative Tajima's D values
160 (Fig. 2c). The distribution of Tajima's D in *C. h. aquapotentis* is significantly smaller than
161 zero (one-sided t-test: t = -28.883, df = 681, p-value < 2.2e-16), which is consistent with
162 recent population expansion. Simulation-based projections of the evolution of the overall *C.*
163 *hominis* population finds it is being shaped by 'recent gene flow', likely indicating recent
164 secondary contact (Fig. 2c-d) between the two subspecies ~165 (24 – 647; 5-95% CI) years
165 ago.

166 We found evidence of two selective sweeps on chromosome 6 in *C. h. hominis* based
167 on composite likelihood ratio (CLR) statistic using SweeD[30] (Supplementary Fig. 1).

168 However, we do not see any other hallmarks of a selective sweep in these regions based on π
169 or Tajima's D (see Supplementary Fig. 2). Possibly, the selective sweeps were incomplete or
170 occurred in the distant past, eroding their genomic signatures. Nevertheless, this may have
171 contributed to the recent decline in effective population size of *C. h. hominis* (Fig. 2a);
172 genetic variation could have been lost during the selective sweep not only in the affected
173 chromosomal regions, but throughout the genome in this largely clonally reproducing
174 organisms, as the selectively favoured variant replaced other existing variants.
175

176 **Linkage, recombination, introgression and gene flow between subspecies**

177 *Distinct patterns of decay of linkage disequilibria*

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

184

185 *Recombination and regions of secondary contact*

186 Using RDP4[31], we identified significant recombination events between the two clades in
187 chromosome 1 (event 1), 2 (event 2) and 6 (event 3 and 4) (Fig; 3b and Supplementary Table
188 2). This indicated secondary contacts between *C. h. hominis* and *C. h. aquapotentis*
189 (IbA10G2), which resulted in rare genetic exchanges between these otherwise diverged
190 clades. We also undertook additional analyses of the highly admixed European isolate

191 (UK_UKH4 of *gp60* subtype IaA14R3) that showed unique ancestry (Fig 1B) and clustered
192 with low-income countries (*C. h. hominis*) (see Supplementary Text).
193

194 *Signature of introgression and gene flow between the clades*

195 We analysed the recombination events in more detail to better understand the implications of
196 genetic introgression between the two subspecies. Determining the signature of genetic
197 introgression is crucial as these regions can also be responsible for increasing genetic
198 diversity and providing novel substrate for natural selection in host-parasite coevolution[32].
199 We used HybridCheck[33] to perform introgression analyses for recombinant events 2 – 4
200 (event 1 was excluded due to a missing parental sequence). We randomly selected a triplet
201 (recombinant, minor parent and major parental lines) from the RDP output (Supplementary
202 Table 2), which revealed a clear signature of introgression (Fig. 3 c-d), also supported by an
203 ABBA-BABA test (Fig. 3e). Additionally, we calculated the pairwise r^2 between SNPs within
204 chromosomes of *C. h. hominis* to assess linkage among SNPs in the introgressed regions (Fig.
205 3f-g). Large blocks of high LD that encompass the introgressed regions suggested each had
206 been exchanged as a single event between *C. h. hominis* and *C. h. aquapotentis* (IbA10G2)
207 ~165 (24 – 647; 5-95% CI) years ago.
208

209 **Could recent introgression increase virulence?**

210 Our analyses suggest *C. h. hominis* and *C. h. aquapotentis* have been diverged and largely
211 reproductively isolated for ~500 years. Noting that the latter subspecies has come to
212 dominate infections within high-income countries[34], is more virulent[25], appears better able to
213 transmit through direct human-to-human contact[24] and possibly at a lower infectious dose[22],
214 *C. h. aquapotentis* may owe its success to being better adapted to human infection. If this is
215 the case, it is possible that recent introgression between these two subspecies could select for

216 more virulent and transmissible *C. hominis* subspecies in low-income countries, particularly
217 noting the recent genetic bottlenecking we have observed in *C. h. hominis* within the last
218 decades.
219
220 *Identification of potential virulence genes*

221 To explore this, we first predicted candidate virulence genes in *C. hominis* likely to be
222 involved in the host-parasite interaction and engaged in a coevolutionary arms race with the
223 host. Broadly, such genes are under continuous adaptation and involved in reciprocal,
224 adaptive genetic changes while interacting with host[32]. During coevolution, evolutionary
225 forces act on virulent genes to create genetic variation through mutation, recombination and
226 gene flow and moulds genetic variation through selection. To identify putative virulence
227 genes in *C. hominis*, we selected the top 5% most highly polymorphic genes based on
228 nucleotide diversity (π). These were filtered for the top 25% of genes under balancing
229 selection, based ranked Tajima's D. Finally, these genes were filtered by selecting the top
230 50% of genes with the highest proportion of non-synonymous mutations, based on ranked
231 Ka/Ks ratio. Using this approach, we identified 24 highly polymorphic genes (Supplementary
232 Table 4) that are rapidly mutating at the protein level and under selective pressure. These
233 genes were significantly more polymorphic (two-sided t-test: t = -2.9062, df = 23, p-value =
234 0.007957), under stronger balancing selection (two-sided t-test: t = -10.393, df = 23, p-value
235 = 2.533e-10) and positive selection (two-sided t-test: t = -2.4736, df = 23, p-value = 0.021)
236 compared to all remaining polymorphic genes. Moreover, these genes are enriched in
237 recombination regions ($\chi2$= 225.04, df = 1, p-value = 0.00049), showing that gene flow
238 within *C. hominis* is inclined towards elevated levels of genetic variation. Overall, these
239 genes are enriched for extracellular ($\chi2$= 13.608, df = 1, p-value = 0.00349) and signal
240 peptide proteins ($\chi2$= 6.69, df = 1, p-value = 0.015), which is consistent with their potential

241 relevance for host-parasite interaction. Together, these results provided the evidence of
242 genomic islands of putative virulence genes (GIPVs: Fig. 4) that are likely to be in a
243 coevolutionary arms race with the host, undergoing frequent recombination and accumulating
244 beneficial polymorphisms that are under selection, and that this increased population
245 diversity.

246

247 *Subspecific divergence of putative virulence genes*

248 We investigated whether any virulence genes predicted above were highly diverged between
249 *C. h. hominis* and *C. h. aquapotentis* and under diversifying selection between the two
250 subspecies. To do this, we calculated absolute divergence (Dxy), representing the average
251 proportion of differences between all pairs of sequences between *C. h. hominis* and *C. h.*
252 *aquapotentis* (IbA10G2), revealing differential gene flow during their reproductive isolation
253 (Fig. 4a). We then calculated the correlation between diversity (π) and divergence (Dxy) for
254 each gene (Fig. 4b). This approach identified four clear outlier genes, which were the most
255 divergent, diverse, and rapidly mutating at the protein level. These genes encoded three
256 mucins (CHUDEA2_430, CHUDEA2_440 and CHUDEA2_450) arrayed in a cluster on
257 chromosome 2, and a hypothetical protein (CHUDEA6_5270) found on chromosome 6 (Fig.
258 4b, Supplementary Table 4). The *gp60* gene ranked 5th in the list of virulence genes but does
259 not sit as a clear outlier from the other genes identified in our assessment. Two of these
260 genes, CHUDEA2_430 (LRT, p-value = 0.005) and CHUDEA6_5270 (LRT, p-value =
261 0.017), are under statistically significant diversifying selection between *C. h. hominis* and *C.*
262 *h. aquapotentis*.

263 Lastly, we looked at each codon position within each of the four outlier genes to
264 detect codon specific diversifying selection, which might be overlooked in the overall gene.
265 We detected episodic diversifying selection (LRT p-value < 0.05) at codon positions 92, 111

266 and 138 of CHUDEA2_430 and 97 and 105 of CHUDEA6_5270 using the Mixed Effects
267 Model of Evolution (MEME) method implemented in Datamonkey[35]. These sites represent
268 putative codons harbouring genetic polymorphisms that experience periods of strong
269 diversifying selection. This strongly suggests CHUDEA2_430 and CHUDEA6_5270 are
270 likely candidate virulence factors participating in a coevolutionary arms race and contributing
271 to the divergence of *C. h. hominis* and *C. h. aquapotentis*.

272

273 *Recent introgression of putative virulence genes*

274 Finally, we asked whether the recent recombination events in chromosomes 2 and 6 between
275 *C. h. hominis* and *C. h. aquapotentis* might include genes associated with increased virulence,
276 indicating increasing virulence of *C. hominis* globally within the past decades, possibly
277 linked to increased human migration. These introgressed regions had significantly elevated
278 levels of nucleotide diversity (two-sided t-test: t = -3.0026, df = 27, p-value = 0.0057),
279 divergence (two-sided t-test = -3.0043, df = 27, p-value = 0.0057) and balancing selection
280 (two-sided t-test: t = -3.0125, df = 27, p-value = 0.0055). Interestingly, we observed a pattern
281 of high diversity and divergence (Fig. 4a), particularly for genes within the recombinant
282 blocks (Fig. 3b-e). The 38 (= top 1%) most divergent genes between *C. h. hominis* and *C. h.*
283 *aquapotentis* were enriched in recombinant regions ($\chi 2=$ 287.08, df = 1, p-value = 0.00049),
284 with 37% present in these regions (Supplementary Table 5), compared to 0.9% of the other
285 1,645 divergent genes. These results suggest the genes undergoing frequent recombination
286 accumulated beneficial polymorphisms that were maintained by balancing selection and that
287 this increased population diversity.

288 Notably, the introgressed regions between *C. h. hominis* and *C. h. aquapotentis*
289 included CHUDEA2_430 (*muc5*) for event 2 (chromosome 2), CHUDEA6_1080 (*gp60*) for

290 event 3 (chromosome 6) and CHUDEA6_5270 (a hypothetical protein) for event 4
291 (chromosome 6). The introgression at CHUDEA6_5270 is particularly intriguing as it sheds
292 further light on the evolution of *C. h. aquapotentis* and *C. h. hominis*, revealing how recent
293 recombination events could be driving the virulence evolution of *C. h. hominis*. We identified
294 two major CHUDEA6_5270 (hypothetical gene) haplotypes; Hap1 representing most *C. h.*
295 *hominis* isolates, and Hap2 representing all *C. h. aquapotentis* plus a small subset of *C. h.*
296 *hominis* (Fig. 5a), as well as many haplotypes associated only with *C. h. hominis*. No
297 mutations were observed in Hap2, which is consistent with *C. h. aquapotentis* having evolved
298 recently, being an estimated 392 years (29 – 1699 years; 5-95% CI) old, assuming a mutation
299 rate of $u=10^{-8}$ and 48h cell division time. Hap1 and Hap2 are diverged by 22 SNPs (Fig. 5b),
300 which is unlikely to represent standing genetic variation, given this is 3-fold higher than the
301 mean deviation among all other CHUDEA6_5270 haplotypes identified here. Instead, we
302 propose CHUDEA6_5270-Hap2 might be an introgressed variant from a highly diverged,
303 unsampled (sub)species that diverged from *C. hominis* around 2.36 million generations ago
304 (assuming a mutation rate of $u=10^{-8}$), which is equal to 12,742 (8,901 – 17,497 years; 5-95%
305 CI) years (assuming 48h/replication). The emergence of CHUDEA6_5270-Hap2 in some *C.*
306 *h. hominis* isolates (i.e., the red section of Hap2 in Fig. 6a) strongly implies that *C. h.*
307 *aquapotentis* has introgressed (circa 400 years ago) into *C. h. hominis* leading to evolution
308 under balancing selection (Fig. 5d).
309 We modelled the 3D protein structure of CHUDEA6_5270-Hap1 and Hap2, and
310 compared these to similarly predicted 3D protein structures for the orthologous genes from *C.*
311 *parvum parvum* and *C. parvum anthropanosum* (Fig. 5c). This modelling identified several
312 conserved alpha-helices that appear to form a coiled-coil domain. CHUDEA6_5270 from *C.*
313 *h. aquapotentis* encodes mutations near the C-terminal end, resulting in a notable kink that
314 deviates from the other structures. This structural variation overlaps with an increase in

315 Tajima's D values toward the C-terminus of the protein, suggesting the region is under
316 balancing selection (Fig. 5d).
317 A similar pattern is observed for CHUDEA2_430 (see Supplementary Fig. 6).
318 However, given that both the subspecies are interspersed in the haplotype network, indicating
319 against a specific directionality of introgression. This may indicate that CHUDEA_430
320 diverged with the divergence of the subspecies and continued to diversify. We were not able
321 to generate a robust 3D structural model for CHUDEA2_430 or its *C. parvum* orthologs. We
322 noted the gene encodes a large intrinsically disordered region (Supplementary Fig. 7) which
323 is a consistent feature of mucin proteins[36], whose structural confirmation is influenced by
324 post-translation glycosylation[37]. Noting this, we did identify eight novel glycosylation sites
325 in *C. h. aquapotentis* CHUDEA2_430 haplotypes not found in *C. h. hominis* (Supplementary
326 Fig 5). Whether these impact interaction with host proteins is not known, but this would be
327 consistent with glycosylated proteins in other pathogens[38].
328

329 **Discussion**

330 We examined the evolutionary genomics of a major human parasite, *C. hominis*, using 114
331 isolates from 16 countries across five continents. We identified two phylogenomic lineages
332 within *C. hominis* that appear largely reproductively isolated, and we propose that these
333 should be considered distinct subspecies. One *C. h. hominis* lineage comprises infections
334 observed in low- to middle-income countries, mostly from Africa and Asia, which we
335 propose be recognised as *C. hominis hominis*, referring to the fact that this subspecies
336 represents the majority population. The second lineage (IbA10G2 subtype) includes isolates
337 from high-income countries, namely Europe, North America and Oceania, which we suggest
338 being named *C. hominis aquapotentis* (strong water), as this subspecies is predominant in
339 countries with longstanding high sanitation and water quality indices.

340 These two subspecies are estimated to have been reproductively isolated for
341 approximately 488 (84 – 2,199) years, except for the more recent genetic exchange at four
342 genetic loci. The reproductive isolation coincides with improvements to sanitation in Europe.
343 It is possible that *C. h. aquapotentis* evolved specialisations making it better suited to human
344 infection allowing it to be more successful through direct transmission supported by a lower
345 infectious dose. Such adaptations would have allowed *C. h. aquapotentis* to become
346 dominant in higher-income countries where sanitation has reduced the level of environmental
347 transmission. Indeed, epidemiological investigations in the UK identified direct person-to-
348 person transmission as a key pathway for *C. h. aquapotentis*[24]. We hypothesise this may have
349 resulted in its rapid population expansion over the last ~500 years and (partial) reproductive
350 isolation. In contrast, *C. h. hominis* experienced recent reduction in effective population size
351 (*Ne*) within the past few decades. We detected two signatures of selective sweep in this
352 subspecies, and we propose that this may have eroded some of the genetic variation, resulting
353 in the marked drop in *Ne* = 5000 to *Ne* = 100 in the past.

354 Despite increased migration and international travelling in the past decades, our study
355 suggests that there has been relatively little movement between continents for this parasite.
356 This is in contrast to reports for *C. parvum.* Corsi *et. al.* recently found a higher proportion of
357 admixture and gene flow between *C. parvum* populations and no evidence of population
358 structuring by geographic region[39]. Despite the strong population sub-structuring in *C.*
359 *hominis*, we found evidence of potential recombination and gene flow between the
360 geographic populations and subspecies. We further investigated and identified the
361 introgressed regions where we detected significant gene flow between the low- and high-
362 income countries. Simulation-based analyses indicated this was most likely explained by
363 'recent geneflow' (circa 165 years ago). This would appear to be a secondary contact
364 between the two subspecies after recent globalisation, illustrating higher migration rate from

365 high-income to low-income countries, which facilitated gene flow, recombination, population
366 admixture and selective sweep.

367 Finally, we have discovered genomic islands of putative virulence genes (GIPVs)
368 contributing to population diversification between *C. h. aquapotentis* and *C. hominis hominis*.
369 These islands have experienced relatively elevated recombination rate which has enriched
370 nucleotide variation under balancing selection and the acquisition of non-synonymous SNPs,
371 consistent with virulence factors driving host-parasite interactions. Intriguingly, the most
372 significant signals within these analyses are driven by *gp60*, a hypothetical protein
373 (CHUDEA6_5270) and a cluster of mucin-like genes (CHUDEA2_430, CHUDEA2_440 and
374 CHUDEA2_450) found on chromosome 2. These genes are consistently identified as being
375 under selection in the evolution of the *C. hominis* subspecies here and in similar observations
376 made of *C. hominis* in Africa[40]. Their orthologs are associated with recombination between
377 human-specific *C. parvum anthroponosum* relative to the zoonotic *C. parvum parvum* and
378 appear to have driven convergence of the former with *C. hominis*[16].

379 CHUDEA2_430 (*muc5*) and hypothetical protein CHUDEA6_5270 are the most
380 notable, displaying significant diversifying selection between the two subspecies. Broadly,
381 mucins mediate cell-cell interactions[41], and modulate infectivity of *Cryptosporidium*
382 sporozoites and merozoites and oocyst production[41,42]. In *C. parvum*, MUC5 is involved in
383 host-cell invasion and an important determinant of host adaptation[41] and highly expressed in
384 the first 2 hours of infection in vitro[43]. MUC5 may also play a role in tethering the sporozoite
385 to the oocyst wall[44]. Our analyses suggest *C. h. aquapotentis* and *C. h. hominis muc5*
386 haplotypes diverged before or with the subspecies and subsequently diversified, which is
387 consistent with prior observations implicating CHUDEA2_430 in the emergence of *C. h.*
388 *aquapotentis*[10,45]. This appears to have resulted in the acquisition of novel glycosylation sites
389 within *C. h. aquapotentis muc5* haplotypes. We cannot determine the functional consequence

390 of these sites but note that glycosylation sites often mediate the specificity of mucin
391 interactions with host proteins in a variety of pathogens[38]. In contrast, CHUDEA6_5270
392 displays a clear signal for the recent introgression of a novel *C. h. aquapotentis* haplotype
393 into *C. h. hominis* after the divergence of these subspecies. This haplotype has notable,
394 structurally relevant, mutations. Identifying the function of this gene, its potential role in
395 infection and the relevance of the structural variation we have inferred here, should be
396 considered a major research priority.

397 In conclusion, this work represents the first large scale population genomic study in
398 any *Cryptosporidium* species, inferring the global population structure and evolutionary
399 history of *C. hominis*. We propose recognition of two distinct subspecies, *C. h. hominis* and
400 *C. h. aquapotentis*, with distinct demographic histories that have diverged circa 500 years
401 ago. Although the subspecies differ in their global distribution, their gene pools are not
402 completely isolated, and rare genetic exchanges have occurred in the recent past. We contend
403 that many of the genes, CHUDEA2_430 and CHUDEA6_5270 in particular, in these
404 introgression regions are involved in infection, and that their evolution in humans may be
405 driving greater human specificity, virulence and transmissibility. It appears *C. h.*
406 *aquapotentis* is playing a key role in this process, which is supported by previous
407 observations based on multilocus typing[28]. This illustrates how human-mediated gene flow is
408 involved in parasite evolution and genomic architecture, and how it could affect virulence
409 evolution. Also, it shows that the GIPVs that result from population admixture in an
410 anthroponotic species are under selection and involved in evolutionary arms race.

411

412 **Methods**

413 **Parasite isolates**

414 The *C. hominis* isolates newly sequenced for this study (n = 34) were archived stool samples
415 collected at the *Cryptosporidium* Reference Unit in the UK. The species was determined by
416 species-specific real-time PCR targeting the A135 gene[46] and subtyped by PCR and
417 sequencing of the *gp60* gene[47]. Supplementary Table 1 provides information about these
418 isolates. Isolates were selected to mainly represent the dominant variant, IbA10G2, as defined
419 by *gp60* sequencing.
420
421 **Processing of faecal samples for whole genome sequencing**
422 Stool samples were processed as previously described[48]. Briefly, saturated salt-flotation was
423 used to obtain a partially purified suspension of oocysts starting from 1-2 ml of each faecal
424 sample. Oocysts were further purified from the suspension by immunomagnetic separation
425 (IMS), using the Isolate® IMS kit (TCS Biosciences, Botolph Claydon, UK). IMS-purified
426 oocysts were treated with bleach, and washed three times with nuclease-free water by
427 centrifugation at 1,100X g for 5 min. The pellets were suspended in 200 µL of nuclease-free
428 water for DNA extraction.
429
430 **DNA preparation and whole genome sequencing**
431 Genomic DNA was extracted from purified *Cryptosporidium* oocysts by first performing
432 eight cycles of freezing in liquid nitrogen for 1 min and thawing at 95°C for 1 min, and then
433 using the QIAamp DNA extraction kit (Qiagen, Manchester, UK) according to the
434 manufacturer's instructions. The genomic DNA was eluted in 50 µL nuclease-free water, and
435 the concentration measured using the Qubit dsDNA HS Assay Kit with the Qubit 1.0
436 fluorometer (Invitrogen, Paisley, UK), according to the manufacturer's instructions.
437

438 Whole genome amplification (WGA) was performed using the Repli-g Midi kit
439 (Qiagen, Milan, Italy), according to the manufacturer's instructions. Briefly, 5 µL of genomic
440 DNA (containing 1-10 ng of DNA) were mixed with 5 µL of denaturing solution and
441 incubated at room temperature for 3 min. Next, 10 µL of stop solution were added to stabilise
442 denatured DNA fragments. The reaction mixture was completed with 29 µL of buffer and 1
443 µL of phi29 polymerase, and allowed to proceed for 16 hours at 30°C. The reaction was
444 stopped by heating at 63°C for 5 minutes. WGA products were visualised by electrophoresis
445 on a 0.7% agarose gel, purified and quantified by Qubit as described above.
446 For Next Generation Sequencing (NGS) experiments, about 1 µg of purified WGA
447 product was used to generate Illumina TruSeq 2x 150 bp paired-end libraries (average insert
448 size: 500 bp), which were sequenced on an Illumina HiSeq 4000 platform (Illumina,
449 SanDiego, CA). Library preparation and NGS experiments were performed by a commercial
450 company (GATC, Germany).
451
452 **Whole genome global dataset**
453 To perform a global comparative genomics of *C. hominis*, we supplemented our newly
454 sequenced genome dataset by downloading all available published *C. hominis* genomes on till
455 date (25th July 2021), from the sequence read archive (SRA) of NCBI and from the EMBL's
456 European Nucleotide Archive (ENA) (see Supplementary Table 1). Collectively, these data
457 represented 114 genomes from 16 countries across five continents.
458
459 **Data pre-processing and variant calling**
460 Raw reads of the 114 *C. hominis* isolates were trimmed to remove adapter sequences and
461 filtered for low-quality bases using Trimmomatic v.0.36[49]. The filtered reads were aligned to

462 *C. hominis* UdeA01 reference genome[50,51] using the maximal exact matches (MEM)
463 algorithm implemented in Burrows-Wheeler Alignment (BWA) tool v.0.7[52] with default
464 settings. Sequence variants were called from the aligned reads of each isolate using the
465 HaplotypeCaller method in the Genome Analysis Toolkit (GATK) v3.7.0[53] as per GATK's
466 best practices pipeline[54]. Called variants were removed if quality depth (QD) < 2.0, Fisher
467 strand (FS) > 60.0, mapping quality (MQ) < 40.0, mapping quality rank sum test
468 (MQRankSum) < -12.5, read position rank sum test (ReadPosRankSum) < -8.0, Strand odds
469 ratio (SOR) > 4.0, QUAL < 30, allele depth (AD) < 5, MAF < 0.05 or missingness > 0.5.
470 Each of the 114 whole genomes assessed here had > 80% coverage of the *C. hominis*
471 reference genome to at least the 5-fold depth. All identified variants were combined in one
472 file and each isolate genotyped using the GenotypeGVCFs tool (GATK v3.7.0) [53] and the
473 variants were separated into SNPs (i.e., nucleotide substitutions) and INDELs.
474
475 **Population genetic structure based on whole genome SNPs**
476 The filtered SNPs were used for population structure, phylogenetic and clustering analyses.
477 Samples having multiple co-infections were identified using estMOI[55] and MOIMIX
478 (https://github.com/bahlolab/moimix). The R package SNPRelate v.1.18[56] was used for
479 principal-component analysis (PCA) analysis. A maximum likelihood phylogenetic tree was
480 constructed by IQ-TREE[57] with 1000 bootstraps and visualised in iTOL v3[58]; the sister
481 species *C. parvum* was used as an outgroup. We also constructed a consensus of 10^7 trees
482 using DensiTree 2[59] in BEAST v2[60]. BEAST v2[60] was also used to estimate the divergence
483 time between the populations by using 95% highest posterior density (HPD); and SpeedDate
484 (https://github.com/vanOosterhoutLab/SpeedDate.jl) to estimate the coalescence times
485 between sequences by using 5-95% confidence interval (CI). We used mutation rate of 10^{-8}
486 and a generation time of 48h/replication[16] to date the coalescence times between sequences.

487 A Neighbor-Net algorithm based network was generated using SplitsTree5[61]. Genetic
488 structure was analysed by STRUCTURE v2.3 software[62] for population number (K) ranging
489 2 - 10 and plotted by using plotSTR R package (https://github.com/DrewWham/Genetic-
490 Structure-Tools). The optimal population genetic cluster value K was estimated by using
491 CLUMPAK[63].
492
493 **Population demographic history and divergence time estimation**
494 We used Bayesian Markov Chain Monte Carlo (MCMC) method implemented in Beast v2
495 program[60] to estimate the effective population size (*Ne*) of the *C. hominis* population. The
496 nucleotide substitution model of HKY was selected. A strict molecular clock model and a
497 Bayesian skyline coalescent tree prior was used with 10^9 generations of MCMC chain and
498 10% burn-ins. Tracer v.1.7[64] was used to assess chain convergence and effective sample size
499 [ESS] > 200 and to construct the demographic history over time; i.e. Bayesian Skyline Plot
500 (BSP). SweeD[30] was used to detect windows of selective sweeps from genome-wide SNP
501 dataset by using composite likelihood ratio (CLR) statistic that identifies signature of site
502 frequency spectrum (SFS), with a grid size of 1000.
503 Demographic histories and migration rates between the *C. hominis* populations were
504 estimated by using fastsimcoal2[65] by using mutation rate of 10^{-8} and a generation time of
505 48h/replication[16]. We first inferred best parameters and the likelihoods for each of the
506 demographic models – no geneflow, ongoing geneflow, early geneflow, recent geneflow and
507 different geneflow, since the time of divergence (~500 years) by running 100 independent
508 iterations with 300,000 coalescent simulations and 60 optimisation cycles. Demographic
509 model with the highest likelihood was then selected to run parameter estimation with block-
510 bootstrapping of 100 replicates.
511

512 **Linkage, recombination and gene flow analyses**

513 We inferred rate of decay of linkage disequilibrium (LD) by calculating the squared
514 correlation of the coefficient (r^2) between SNPs within 50kb using VCFtools[66]. LD blocks
515 were also determined by calculating pairwise r^2 between SNPs within chromosomes of each
516 population. Recombination events were identified using the Recombination Detection
517 Program version 4 (RDP4)[31] using the RDP[67], GENECONV[68], BootScan[69], MaxChi[70] and
518 Chimaera[71] methods. Events were considered significant if at least three methods predicted
519 their occurrence at a probability values, $p \leq 10^{-5}$. Recombination events with undetermined
520 parental sequences were excluded from further HybridCheck analyses. Statistically
521 significant recombination events were visualised and analysed using HybridCheck[33] to
522 determine the sequence similarity between the isolates involved in the events. HybridCheck
523 program was also used to calculate the D statistic and estimate the gene flow between the
524 populations.

525

526 **Population genetic and genomic analyses of coding region**

527 Tajima's D, Nucleotide diversity (π), Dxy and Fst were calculated using the PopGenome R-
528 package[72]. Nonsynonymous (Ka) and synonymous (Ks) mutation rates were calculated by
529 using Ka/Ks_Calculator[73]. Protein localisation (extracellular) was predicted using WoLF
530 PSORT[74] and information regarding predicted protein targeting (signalling peptides) genes
531 were obtained from CryptoDB[50]. POPART program was used to generate haplotype
532 networks[75]. AlphaFold was used to predict the protein structures[76]. Glycosylation sites were
533 predicted by using NetNGlyc 4.0 Server[77]. Intrinsically disordered region in proteins were
534 predicted using IUPred2A[78]. All statistical tests and results were performed and plotted in R
535 (version 3.6.1).

536

537 **Data availability**

538 All raw and processed sequencing data generated and analysed during the current study have
539 been submitted to the NCBI Sequence Read Archive
540 https://www.ncbi.nlm.nih.gov/bioproject/), under the BioProject PRJEB15112,
541 PRJNA610731, PRJNA610732, PRJNA610735, PRJNA610737, PRJNA610738,
542 PRJNA610739, PRJNA610740, PRJNA610741, PRJNA610742, PRJNA610743,
543 PRJNA610744, PRJNA610745, PRJNA610746, PRJNA610747 and PRJNA610748.

544

545 **References**

546 1. Khalil, I.A., *et al.* Morbidity, mortality, and long-term consequences associated with
547 diarrhoea from Cryptosporidium infection in children younger than 5 years: a meta-
548 analyses study. *The Lancet. Global health* **6**, e758-e768 (2018).
549 2. Kotloff, K.L., *et al.* Burden and aetiology of diarrhoeal disease in infants and young
550 children in developing countries (the Global Enteric Multicenter Study, GEMS): a
551 prospective, case-control study. *Lancet (London, England)* **382**, 209-222 (2013).
552 3. Ryan, U., Zahedi, A. & Paparini, A. Cryptosporidium in humans and animals—a one
553 health approach to prophylaxis. *Parasite immunology* **38**, 535-547 (2016).
554 4. Santin, M. Cryptosporidium and Giardia in Ruminants. *Vet Clin North Am Food Anim*
555 *Pract* **36**, 223-238 (2020).
556 5. Razakandrainibe, R., *et al.* Common occurrence of Cryptosporidium hominis in
557 asymptomatic and symptomatic calves in France. *PLoS neglected tropical diseases* **12**,
558 e0006355 (2018).
559 6. Yang, X., Guo, Y., Xiao, L. & Feng, Y. Molecular epidemiology of human
560 cryptosporidiosis in low-and middle-income countries. *Clinical Microbiology Reviews*
561 **34**, e00087-00019 (2021).
562 7. Putignani, L. & Menichella, D. Global distribution, public health and clinical impact of
563 the protozoan pathogen cryptosporidium. *Interdisciplinary perspectives on infectious*
564 *diseases* **2010**(2010).
565 8. Craun, G.F., Hubbs, S.A., Frost, F., Calderon, R.L. & Via, S.H. Waterborne outbreaks of
566 cryptosporidiosis. *Journal AWWA* **90**, 81-91 (1998).
567 9. Nichols, G.L., Chalmers, R.M. & Hadfield, S.J. Molecular epidemiology of human
568 cryptosporidiosis. in *Cryptosporidium: parasite and disease* 81-147 (Springer, 2014).
569 10. Feng, Y., Ryan, U.M. & Xiao, L. Genetic Diversity and Population Structure of
570 Cryptosporidium. *Trends in parasitology* (2018).
571 11. Herges, G.R., *et al.* Evidence that Cryptosporidium parvum populations are panmictic
572 and unstructured in the Upper Midwest of the United States. *Applied and*
573 *environmental microbiology* **78**, 8096-8101 (2012).

12. Wang, R.*, et al.* Cryptosporidium parvum IId family: clonal population and dispersal from Western Asia to other geographical regions. *Scientific reports* **4**, 4208 (2014).
13. Morrison, L.J.*, et al.* The population structure of the Cryptosporidium parvum population in Scotland: a complex picture. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **8**, 121-129 (2008).
14. King, P.*, et al.* Prevalence and epidemiology of human Cryptosporidium parvum IIc infections in England and Wales. *The Lancet* **389**, S56 (2017).
15. King, P., Tyler, K.M. & Hunter, P.R. Anthroponotic transmission of Cryptosporidium parvum predominates in countries with poorer sanitation: a systematic review and meta-analysis. *Parasites & vectors* **12**, 16-16 (2019).
16. Nader, J.L.*, et al.* Evolutionary genomics of anthroponosis in Cryptosporidium. *Nature Microbiology* (2019).
17. Yang, X., Guo, Y., Xiao, L. & Feng, Y. Molecular Epidemiology of Human Cryptosporidiosis in Low- and Middle-Income Countries. *Clinical microbiology reviews* **34**, e00087-00019 (2021).
18. Jex, A.R. & Gasser, R.B. 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. *Biotechnology advances* **28**, 17-26 (2010).
19. Efstratiou, A., Ongerth, J.E. & Karanis, P. Waterborne transmission of protozoan parasites: review of worldwide outbreaks-an update 2011–2016. *Water research* **114**, 14-22 (2017).
20. Chalmers, R.M.*, et al.* 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 (2010).
21. Widerström, M.*, et al.* Large outbreak of Cryptosporidium hominis infection transmitted through the public water supply, Sweden. *Emerg Infect Dis* **20**, 581 (2014).
22. Segura, R., Prim, N., Montemayor, M., Valls, M.E. & Muñoz, C. Predominant virulent IbA10G2 subtype of Cryptosporidium hominis in human isolates in Barcelona: a five-year study. *PloS one* **10**, e0121753 (2015).
23. Zhou, L., Singh, A., Jiang, J. & Xiao, L. Molecular surveillance of Cryptosporidium spp. in raw wastewater in Milwaukee: implications for understanding outbreak occurrence and transmission dynamics. *Journal of clinical microbiology* **41**, 5254-5257 (2003).
24. McKerr, C.*, et al.* Cross-Sectional Household Transmission Study of Cryptosporidium Shows that C. Hominis Infections are A Key Risk Factor for Spread. (2021).
25. Cama, V.A.*, et al.* Cryptosporidium species and subtypes and clinical manifestations in children, Peru. *Emerg Infect Dis* **14**, 1567 (2008).
26. Cacciò, S.M. & Chalmers, R.M. Human cryptosporidiosis in Europe. *Clin Microbiol Infect* **22**, 471-480 (2016).
27. Feng, Y., Tiao, N., Li, N., Hlavsa, M. & Xiao, L. Multilocus sequence typing of an emerging Cryptosporidium hominis subtype in the United States. *Journal of clinical microbiology* **52**, 524-530 (2014).
28. Li, N.*, et al.* Genetic recombination and Cryptosporidium hominis virulent subtype IbA10G2. *Emerging infectious diseases* **19**, 1573-1582 (2013).
29. Drummond, A.J., Rambaut, A., Shapiro, B. & Pybus, O.G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution* **22**, 1185-1192 (2005).

621 30. Pavlidis, P., Živkovic, D., Stamatakis, A. & Alachiotis, N. SweeD: likelihood-based
622 detection of selective sweeps in thousands of genomes. *Molecular biology and*
623 *evolution* **30**, 2224-2234 (2013).
624 31. Martin, D.P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and
625 analysis of recombination patterns in virus genomes. *Virus Evolution* **1**(2015).
626 32. Van Oosterhout, C. Mitigating the threat of emerging infectious diseases; a
627 coevolutionary perspective. *Virulence* **12**, 1288-1295 (2021).
628 33. Ward, B.J. & van Oosterhout, C. HYBRIDCHECK: software for the rapid detection,
629 visualization and dating of recombinant regions in genome sequence data. *Mol Ecol*
630 *Resour* **16**, 534-539 (2016).
631 34. Guo, Y.*, et al.* Comparative genomic analysis reveals occurrence of genetic
632 recombination in virulent Cryptosporidium hominis subtypes and telomeric gene
633 duplications in Cryptosporidium parvum. *BMC genomics* **16**, 320 (2015).
634 35. Pond, S.L. & Frost, S.D. Datamonkey: rapid detection of selective pressure on
635 individual sites of codon alignments. *Bioinformatics (Oxford, England)* **21**, 2531-2533
636 (2005).
637 36. Carmicheal, J.*, et al.* Presence and structure-activity relationship of intrinsically
638 disordered regions across mucins. *The FASEB Journal* **34**, 1939-1957 (2020).
639 37. Perez-Vilar, J. & Hill, R.L. The Structure and Assembly of Secreted Mucins* 210. *Journal*
640 *of Biological Chemistry* **274**, 31751-31754 (1999).
641 38. Lin, B., Qing, X., Liao, J. & Zhuo, K. Role of Protein Glycosylation in Host-Pathogen
642 Interaction. *Cells* **9**, 1022 (2020).
643 39. Corsi, G.I.*, et al.* Evolutionary epidemiology of a zoonosis. *bioRxiv*,
644 2021.2010.2015.464618 (2021).
645 40. Tichkule, S.*, et al.* Comparative genomics revealed adaptive admixture in
646 Cryptosporidium hominis in Africa. *Microb Genom* **7**(2021).
647 41. O'Connor, R.M.*, et al.* Polymorphic mucin antigens CpMuc4 and CpMuc5 are integral
648 to Cryptosporidium parvum infection in vitro. *Eukaryot Cell* **8**, 461-469 (2009).
649 42. Cevallos, A.M.*, et al.* Mediation of Cryptosporidium parvum infection in vitro by
650 mucin-like glycoproteins defined by a neutralizing monoclonal antibody. *Infection and*
651 *immunity* **68**, 5167-5175 (2000).
652 43. Lippuner, C.*, et al.* RNA-Seq analysis during the life cycle of Cryptosporidium parvum
653 reveals significant differential gene expression between proliferating stages in the
654 intestine and infectious sporozoites. *International journal for parasitology* **48**, 413-422
655 (2018).
656 44. Chatterjee, A.*, et al.* Evidence for mucin-like glycoproteins that tether sporozoites of
657 Cryptosporidium parvum to the inner surface of the oocyst wall. *Eukaryot Cell* **9**, 84-
658 96 (2010).
659 45. Bouzid, M., Hunter, P.R., Chalmers, R.M. & Tyler, K.M. Cryptosporidium pathogenicity
660 and virulence. *Clin Microbiol Rev* **26**, 115-134 (2013).
661 46. Robinson, G., Elwin, K. & Chalmers, R.M. Cryptosporidium Diagnostic Assays:
662 Molecular Detection. *Methods in molecular biology (Clifton, N.J.)* **2052**, 11-22 (2020).
663 47. Chalmers, R.M., Robinson, G., Elwin, K. & Elson, R. Analysis of the Cryptosporidium
664 spp. and gp60 subtypes linked to human outbreaks of cryptosporidiosis in England and
665 Wales, 2009 to 2017. *Parasites & Vectors* **12**, 95 (2019).

666 48. Hadfield, S.J., *et al.* Generation of whole genome sequences of new Cryptosporidium
667 hominis and Cryptosporidium parvum isolates directly from stool samples. *BMC*
668 *genomics* **16**, 650 (2015).
669 49. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
670 sequence data. *Bioinformatics (Oxford, England)* **30**, 2114-2120 (2014).
671 50. Heiges, M., *et al.* CryptoDB: a Cryptosporidium bioinformatics resource update.
672 *Nucleic acids research* **34**, D419-422 (2006).
673 51. Isaza, J.P., *et al.* Revisiting the reference genomes of human pathogenic
674 Cryptosporidium species: reannotation of C. parvum Iowa and a new C. hominis
675 reference. *Scientific reports* **5**, 16324 (2015).
676 52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
677 transform. *Bioinformatics (Oxford, England)* **25**, 1754-1760 (2009).
678 53. McKenna, A., *et al.* The Genome Analysis Toolkit: a MapReduce framework for
679 analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303
680 (2010).
681 54. Van der Auwera, G.A., *et al.* From FastQ data to high confidence variant calls: the
682 Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics*
683 **43**, 11.10.11-33 (2013).
684 55. Assefa, S.A., *et al.* estMOI: estimating multiplicity of infection using parasite deep
685 sequencing data. *Bioinformatics (Oxford, England)* **30**, 1292-1294 (2014).
686 56. Zheng, X., *et al.* A high-performance computing toolset for relatedness and principal
687 component analysis of SNP data. *Bioinformatics (Oxford, England)* **28**, 3326-3328
688 (2012).
689 57. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: A Fast and
690 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.
691 *Molecular biology and evolution* **32**, 268-274 (2014).
692 58. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and
693 annotation of phylogenetic and other trees. *Nucleic acids research* **44**, W242-245
694 (2016).
695 59. Bouckaert, R.R. & Heled, J. DensiTree 2: Seeing Trees Through the Forest. *bioRxiv*,
696 012401 (2014).
697 60. Bouckaert, R., *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis.
698 *PLOS Computational Biology* **10**, e1003537 (2014).
699 61. Huson, D.H. & Bryant, D. Application of phylogenetic networks in evolutionary studies.
700 *Molecular biology and evolution* **23**, 254-267 (2006).
701 62. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using
702 multilocus genotype data. *Genetics* **155**, 945-959 (2000).
703 63. Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A. & Mayrose, I. Clumpak: a
704 program for identifying clustering modes and packaging population structure
705 inferences across K. *Molecular ecology resources* **15**, 1179-1191 (2015).
706 64. Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. Posterior
707 Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic biology* **67**,
708 901-904 (2018).
709 65. Excofffier, L., *et al.* fastsimcoal2: demographic inference under complex evolutionary
710 scenarios. *Bioinformatics (Oxford, England)* (2021).
711 66. Danecek, P., *et al.* The variant call format and VCFtools. *Bioinformatics (Oxford,*
712 *England)* **27**, 2156-2158 (2011).

67. Martin, D. & Rybicki, E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics (Oxford, England)* **16**, 562-563 (2000).
68. Sawyer, S. GENECONV: a computer package for the statistical detection of gene conversion. *http://www. math. wustl. edu/~ sawyer* (1999).
69. Salminen, M.O., CARR, J.K., BURKE, D.S. & McCUTCHAN, F.E. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS research and human retroviruses* **11**, 1423-1425 (1995).
70. Smith, J.M. Analyzing the mosaic structure of genes. *Journal of molecular evolution* **34**, 126-129 (1992).
71. Posada, D. & Crandall, K.A. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proceedings of the National Academy of Sciences* **98**, 13757 (2001).
72. Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E. & Lercher, M.J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Molecular biology and evolution* **31**, 1929-1936 (2014).
73. Zhang, Z.*, et al.* KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259-263 (2006).
74. Horton, P.*, et al.* WoLF PSORT: protein localization predictor. *Nucleic acids research* **35**, W585-W587 (2007).
75. Leigh, J.W. & Bryant, D. popart: full-feature software for haplotype network construction. *Methods in Ecology and Evolution* **6**, 1110-1116 (2015).
76. Jumper, J.*, et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* (2021).
77. Gupta, R. & Brunak, S. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac Symp Biocomput*, 310-322 (2002).
78. Mészáros, B., Erdős, G. & Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic acids research* **46**, W329-W337 (2018).
79. Lighten, J.*, et al.* Evolutionary genetics of immunological supertypes reveals two faces of the Red Queen. *Nature Communications* **8**, 1294 (2017).

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.

Author contributions

752 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.
753 designed the analyses. S.M.C., R.M.C., G.R., D.E., and K.M.T. were involved in acquisition
754 of data. S.T. performed the bioinformatics associated evolutionary genetic and genomic
755 analyses. A.R.J., S.M.C., C.V.O. and S.T. wrote the manuscript. All authors read and
756 approved the submission of the manuscript for the publication.
757

758 **ORCID ID**

759 Swapnil Tichkule: https://orcid.org/0000-0003-2940-9961

760 Cock van Oosterhout: https://orcid.org/0000-0002-5653-738X
761 Simone M Cacciò: https://orcid.org/0000-0002-8561-1323
762 Samantha J Emery-Corbin: https://orcid.org/0000-0003-2984-0620
763 Kevin M. Tyler: https://orcid.org/0000-0002-0647-8158
764 Guy Robinson: https://orcid.org/0000-0002-6594-2408
765

766 **Competing interests**
767 The authors declare that there are no conflicts of interest.
768

769 **Funding information**
770 A. R. J. acknowledges financial support from the Australian National Health and Medical
771 Research Council (APP1194330) and Victorian State Government Operational Infrastructure
772 Support and Australian Government National Health and Medical Research Council
773 Independent Research Institute Infrastructure Support Scheme. K.M.T. was supported by
774 funding from the Interreg 2 Seas programme 2014–2020 co-funded by the European
775 Regional Development Fund under subsidy contract No. 2S05–043 H4DC. S.M.C
776 acknowledges support from the European Union's Horizon 2020 research and innovation

777 programme, project "Collaborative management platform for detection and analyses of (re-)
778 emerging and foodborne outbreaks in Europe" (COMPARE, www.compare-europe.eu), grant
779 agreement No. 643476. S.T. was supported by Walter and Eliza Hall International PhD
780 Scholarship and Melbourne Research Scholarship (MRS). C.V.O. was supported by the Earth
781 and Life Systems Alliance (ELSA) of the Norwich Research Park (NRP).

782

783 **Figures**

784 **Fig 1.** Global population structure of *C. hominis* isolates illustrating their sub-structing and
785 diversification. **a**. PCA of isolates based on their whole genome SNPs, highlighting three
786 clusters of isolates which are predominately based on continents of origin. Isolates were color
787 coded with their continent of origin. Isolates associated with *gp60* subtype IbA10G2 were
788 represented with solid circles while non-IbA10G2 with solid triangles. **b.** Structure plot
789 illustrating population genetic ancestry and the admixed nature of the *C. hominis* isolates.
790 The plot was obtained assuming an optimum value of K=4. The black arrow (bottom)
791 indicates the highly admixed isolate (UK_UKH4), which includes all four ancestries. **c.**
792 Phylogenetic clustering demonstrating two clades. **d.** Splitstree and **e.** Densitree are also
793 demonstrating two major clades comprising isolates, which are separated based on low- and
794 high-income countries of origin, respectively. *Cryptosporidium h. aquapotentis* includes
795 isolates associated with *gp60* subtype IbA10G2 while *C. h. hominis* includes isolates
796 associated with other *gp60* subtypes.

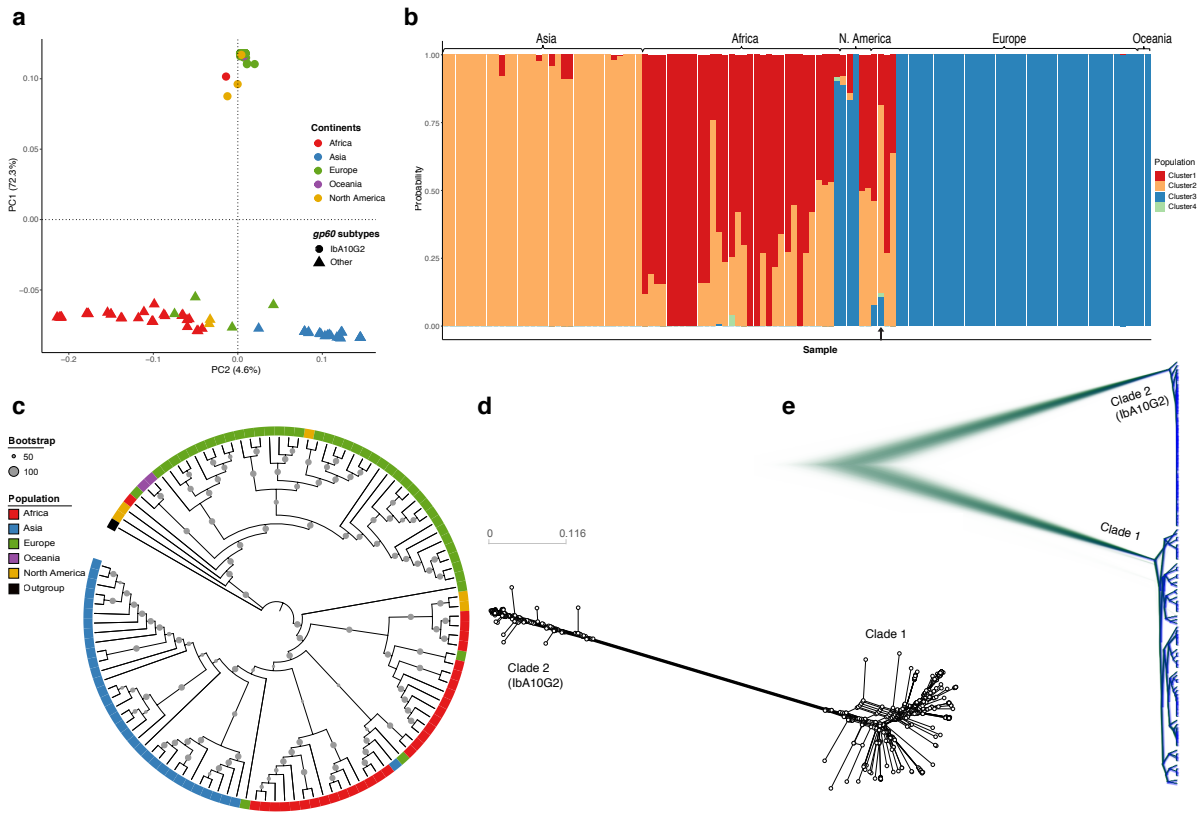


Fig 2. Demographic histories and population size and secondary contact between *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2). **a**. Bayesian Skyline plots (BSP) depicting change in *Ne* (effective population size) through time, for both the clades. The central dark line and the upper 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 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).

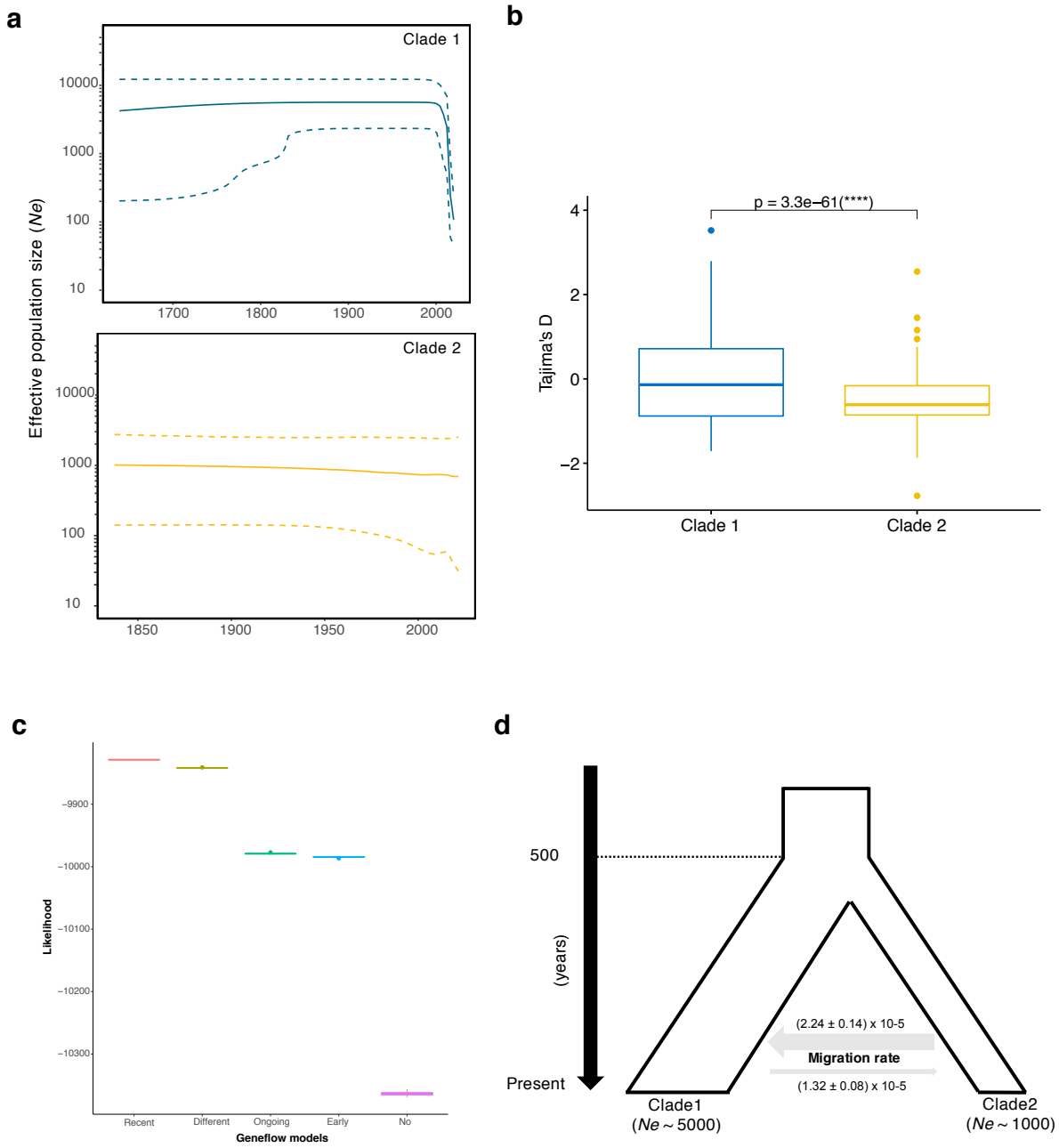
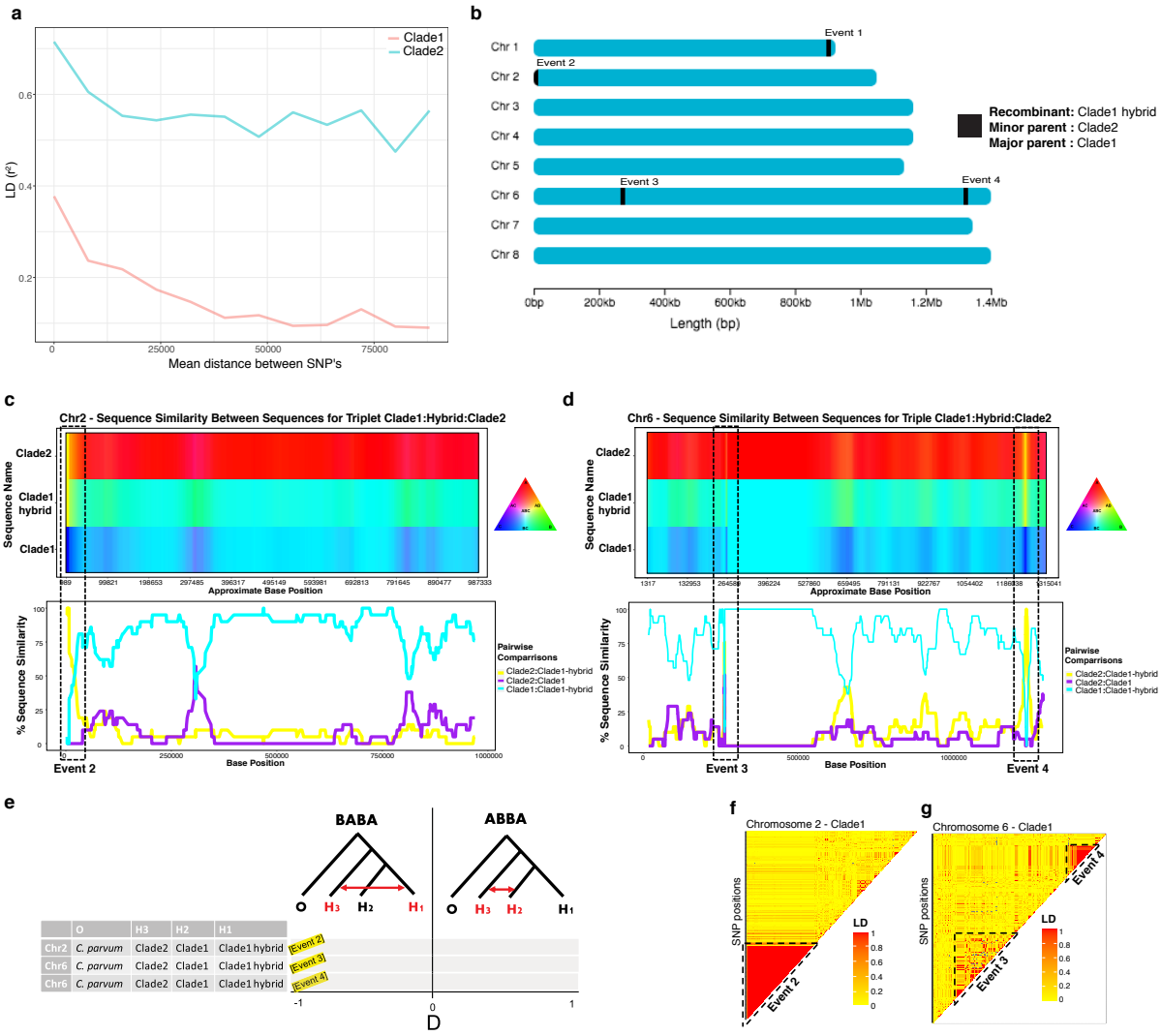


Fig 3. Analyses of recombination and gene flow between *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2). **a**. Linkage disequilibrium (LD) decay plot showing rapid decay of linkage between SNPs in *C. h. hominis* (clade 1) compared to *C. h. aquapotentis* (clade 2). **b**. Graphical representation 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

818 were generated for random set of triplets that includes recombinant (hybrid), minor (donor)
819 and major (recipient) parental sequence, as detected by RDP4 program. Introgressed blocks
820 (recombinant breakpoints) were illustrated with dashed boxes, showing high similarity
821 between the recombinant (*C. h. hominis* hybrid isolates) and minor parent (*C. h. aquapotentis*
822 isolates). **e.** Gene flow analyses with ABBA-BABA test, representing D statistics for the
823 random sets of triplets (as used in c-d) along with *C. parvum* as an outgroup. D statistic
824 values close to -1 at all three recombinant events, suggesting geneflow between H1 and H3.
825 **f-g**. Pairwise LD of SNPs in chromosomes 2 and 6 of *C. h. hominis* showing red-blocks of
826 high linkage between SNPs in introgressed events 2-4.



	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

827

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

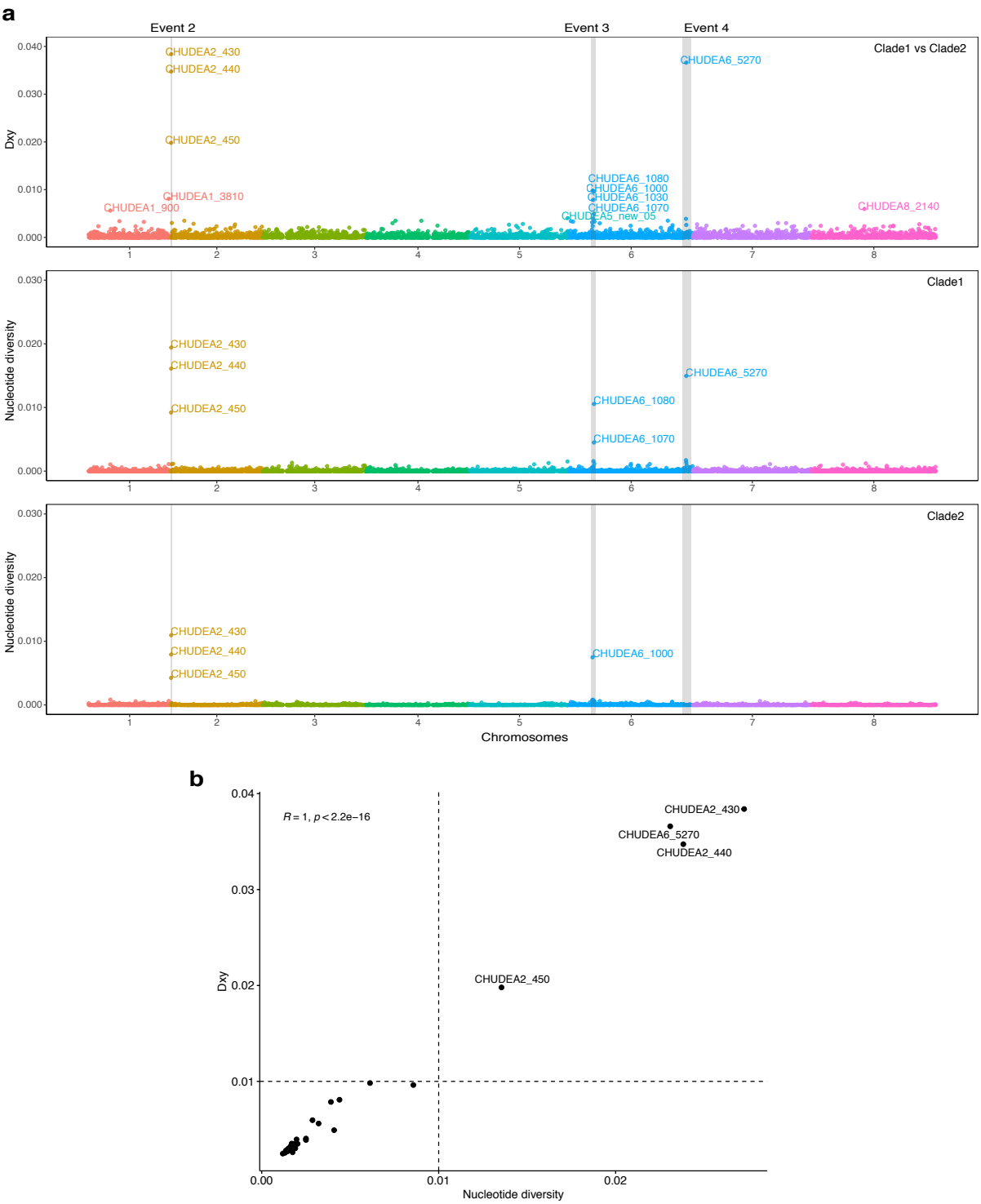
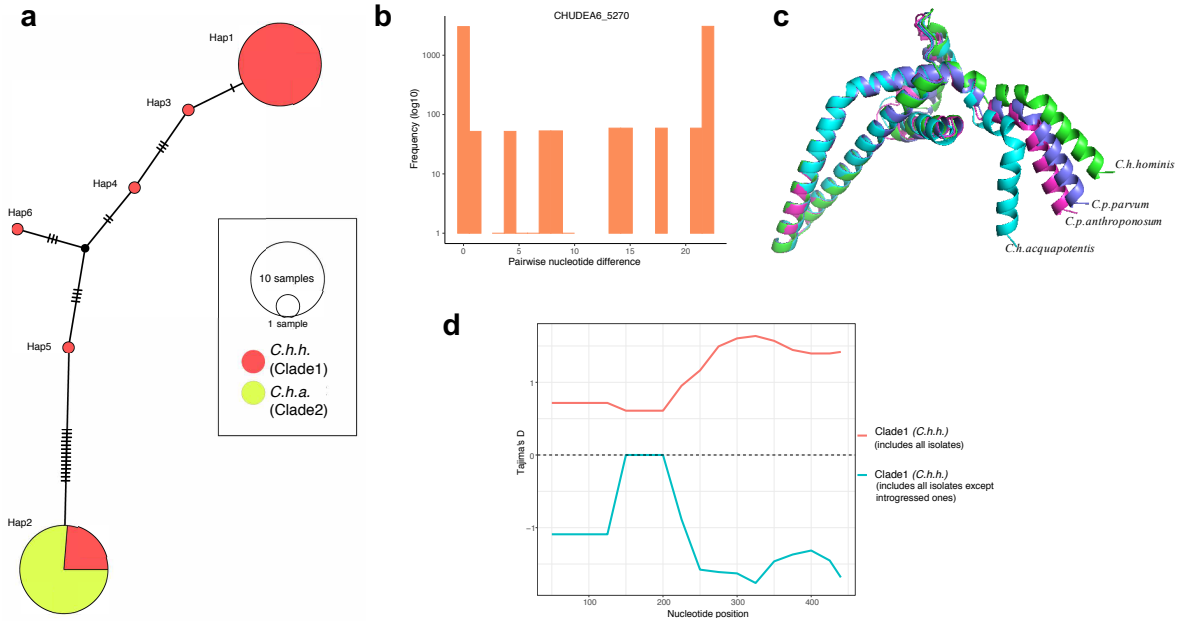


Fig 5. Illustrating diversifying selection between *C. hominis* subspecies and host adaptation at CHUDEA6_5270 (hypothetical gene). **a.** Haplotype network analyses illustrating haplotype diversification between *C. h. hominis* (clade 1) and *C. h. aquapotentis* (clade 2). **b.** Pairwise nucleotide divergence shows bimodal distribution, which, theoretically, can be

842 explained both by balancing selection[79], as well as by genetic introgression. **c.** Comparison of
843 protein structure of CHUDEA6_5270 gene between *Cryptosporidium* species and subtypes
844 demonstrates variation towards C.-terminal region. **d**. Introgressed-isolates driving balancing
845 selection at gene CHUDEA6_5270 in *C. h. hominis*. Red line represents balancing selection
846 (positive Tajima's D) in *C. h. hominis* that also includes introgressed-isolates. Blue line
847 represents purifying selection (negative Tajima's D) in *C. h. hominis* after excluding
848 introgressed-isolates.



849