Title

Population genomics of *Cryptosporidium hominis* across five continents identifies two subspecies that have diverged and recombined 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; genetic introgression; gene flow; secondary contact; selective sweep; evolution.

Repositories

All raw and processed sequencing data generated and analysed during the current study have been submitted to the NCBI Sequence Read Archive https://www.ncbi.nlm.nih.gov/bioproject/), under the BioProject PRJEB15112 and PRJNA607931.

Abstract

Cryptosporidium is a significant public health problem and one of the primary causes of diarrhoea in humans, particularly in very young children living in low- and middle-income countries. While the zoonotic *Cryptosporidium parvum* and anthroponotic *C. hominis* species collectively account for most cases globally, the latter is predominant in low- and middle-income countries. Here, we present a comprehensive whole genome study of *C. hominis,* comprising 114 isolates from 16 countries within five continents. We detect two highly diverged lineages with a distinct biology and demography that have diverged circa 500 years ago. We consider these lineages as two subspecies, and provisionally propose the names *C. hominis hominis* (clade 1) and *C. hominis aquapotentis* (clade 2 or *gp60* subtype IbA10G2). *C. h. hominis* is mostly found in low-income countries in Africa and Asia, and it appears to have recently undergone population contraction. In marked contrast, *C. h. aquapotentis* was found in high-income countries, mainly in Europe, North America and Oceania, and we reveal a signature of population expansion. Moreover, we detected genomic regions of introgression representing gene flow after a secondary contact between the subspecies from low- and high-income countries. We demonstrate that this gene flow resulted in genomic island of high diversity and divergence, and that this diversity at potential virulence genes is maintained by balancing selection, suggesting that they are involved in a coevolutionary arms race.

Introduction

Cryptosporidiosis is one of the leading causes of death due to diarrhoea in children under five globally[1,2], resulting in an estimated 48,000 deaths annually. It is second only to malaria as a parasite contributing to global health burdens, with an overall global disease burden of ~12.8 million Disability Adjusted Life-Years (DALYs)[1]. In humans, cryptosporidiosis is primarily

86 caused by both *Cryptosporidium parvum*, which is zoonotic and a common parasite in young
87 ruminants[3,4], and *C. hominis*, which is likely exclusively anthroponotic and, the more prevalent
88 of the two species globally[5]. The disease burden caused by these parasites is overwhelmingly
89 skewed to low and middle-income countries, particularly in sub-Saharan Africa[1]. However,
90 *Cryptosporidium* parasites are ubiquitous and often associated with large waterborne outbreaks
91 through water treatment paucity or failures in swimming pools and municipal water supplies[6,7].
92 Presently, there are no vaccines or widely available, effective drugs to treat the infection.
93 Hence, control is largely based on prevention of infection, which is underpinned by a strong
94 understanding of the parasite's epidemiology and transmission.

95 Molecular typing has remained the primary method to study the global epidemiology
96 of these parasites, at both a single locus (particularly the highly polymorphic *gp60* gene) or
97 multiple loci (including simple sequence repeats)[8]. This accumulated molecular data has
98 demonstrated significant genetic variability within each species and indicated a complex
99 population genetic structuring. In *C. parvum*, the population structure varies globally from
100 clonal, to epidemic, to panmictic, likely as a result of varying ecologic factors[9-11]. Importantly,
101 specific variants found exclusively with human infections credit the existence of an
102 anthroponotic *C. parvum* lineage (IIc)[12,13], with a recent genome wide study providing strong
103 evidence for two subspecies, the zoonotic *C. parvum parvum* and the anthroponotic *C. parvum*
104 *anthroponosum*[14] . Interestingly, this study also found that both subspecies occasionally still
105 hybridise and exchange genetic variation. These exchanges overlapped with similar genomic
106 regions undergoing genetic introgression between *C. hominis* and *C. parvum anthroponosum*,
107 indicating candidate sites underpinning adaptation to human-specific infection[14].

108 Most molecular studies of *C. hominis* suggest this species is characterised by a largely
109 clonal population structure, with dominance of specific variants in given areas[15]. In particular,
110 the IbA10G2 variant (or subtype) identified from *gp60* sequence appears to be dominant in

111 high-income countries, where it has been associated with large waterborne outbreaks (drinking
112 water and swimming pools) and person-to-person transmission, although it is also found in
113 low- and middle-income countries[16]. Intermixing of genetic material (i.e., recombination)
114 within *C. hominis* appears to be common. A population genetic study of *C. hominis* isolates
115 from Peru suggested that recombination in a region of chromosome 6 encompassing the *gp60*
116 gene may have driven emergence of the IbA10G2 subtype[17]. Relatedly, a recent genomic study
117 of *C. hominis* infecting children in Bangladesh indicated that the high transmission rate and
118 likelihood of mixed infections may be driving higher recombination levels within this
119 species[18]. Moreover, a study in Africa also highlighted evidence of geographically isolated
120 lineages with limited population admixture, which were also likely driven by the exchange of
121 genetic information through recombination[19].

122 To better understand the epidemiology and evolution of *C. hominis*, we aim to address
123 three questions: (1) are the two subtypes identified by *gp60* genotyping (i.e., IbA10G2, and
124 "not-IbA10G2" subtypes) significantly diverged across their entire genome, (2) do they
125 represent reproductively isolated gene pools that no longer exchange genetic information
126 through recombination, and (3) do they possess genes that are significantly diverged as the
127 result of positive or balancing selection? Population genetic analysis of whole genome
128 sequence data from samples collected across five continents enables us to address these
129 questions. The insights gained from these analyses are particularly relevant for public health;
130 the IbA10G2 subtype has already been identified as an emerging host-adapted, virulent, and
131 hyper-transmissible population, making it a threat to human health in high-income
132 countries[17,20,21]. We therefore performed a global study of 114 *C. hominis* genomes, comparing
133 the IbA10G2 subtype to published genomes representing locally acquired infections from five
134 continents. We show that although both subtypes (i.e., IbA10G2, and "not-IbA10G2") are
135 significantly diverged, they continue to exchange a small amount of genetic information

136 through recombination. The genetically introgressed regions comprise genomic islands with
137 drug targets, and the diversity at these potential virulence genes is maintained by balancing
138 selection. Furthermore, the subtypes are estimated to have diverged circa 500 years ago, and
139 the demographic histories of both parasite subtypes are markedly different. We hypothesise
140 that due to the improved water quality and sanitation in high-income countries, the selection
141 pressure on transmission of subtype IbA10G2 has shifted, and that this subtype evolved
142 improved transmissibility through direct human-to-human contact. Given the significant
143 genomic divergence and partial reproductive isolation, and the fact that the transmission of this
144 subtype no longer is strictly dependent on foul water, we propose to name the IbA10G2 subtype
145 *Cryptosporidium hominis aquapotentis* (strong water). Furthermore, we propose to rename the
146 subtype most commonly found in low-income countries in Africa and Asia *Cryptosporidium*
147 *hominis hominis*.
148

149 **Results**

150 **Genomic evidence of population sub-structuring at the continental level**

151 We first tested for the presence of mixed populations in the whole genome sequence data of
152 each of the 114 *C. hominis* isolates but found no evidence for mixed infections (estimated
153 MOI=1). We then identified 5,618 biallelic SNPs which were used to determine the
154 population structure of *C. hominis*. A Principle Component Analysis (PCA) identified three
155 major clusters (Fig 1A). The first component (PC1; 72.3% of the population variation)
156 separated European, North American and Oceania from Asian and African samples, and
157 illustrated that most isolates clustered geographically (107 out of 114, ~94%). The second
158 component (4.6% of the population variation) separated African from Asian samples. Some
159 samples from North America (n=2), Europe (n=4) and Africa (n=1) did not cluster
160 geographically, however, at least some of these represent returned travellers whose infection

161 locations were outside their sampling location. For instance, two European samples, UKH30
162 (from UK) and Sweden_SRR3098109 (from Sweden) reported international travel before
163 sampling to Egypt and Denmark, respectively.

164 The geographical structuring was also evident in a STRUCTURE graph (Fig 1B). 110
165 *C. hominis* isolates are comprised from the three commonly shared population ancestry
166 (clusters 1-3 in Fig 1B), while four samples (three from Africa and one from Europe)
167 possessed a fourth, unique population ancestry (cluster 4 in Fig 1B). Among these four, only
168 one European isolate (UK_UKH4) was assigned to four different populations.

169

170 **Genomic evidence of population diversification**

171 To further support the genetic relationship between *C. hominis* isolates and determine global
172 population structure we generated a Maximum Likelihood (ML) phylogenetic tree (Fig. 1C).
173 We identified two major clades, one corresponding to low-income countries from Asia and
174 Africa (Clade 1), and the other to high-income countries from Europe, North America and
175 Oceania (Clade 2). Clade 2 represents subtype IbA10G2. The divergence between the two
176 clades was estimated to be 488 [84 – 2,199] years ago. The diversification within the global
177 *C. hominis* population was also supported by a network analysis (Fig 1D) and a consensus
178 tree (Fig 1E) generated by SplitsTree and DensiTree, respectively.

179

180 **Demographic histories and selective sweep in Europe**

181 To understand the demographic histories and estimate the change in the effective population
182 size (*Ne*) through time, we constructed a Bayesian Skyline Plot (BSP) [22] for both clades
183 identified in Figure 1. Fig. 2A shows that the parasite isolates from low-income countries

184 (clade 1) experienced a marked population contraction recently from *Ne*=~5000 to *Ne*=~100,
185 which was supported by a higher proportion of positive Tajima's D values (Fig 2B). In
186 contrast, the isolates from high-income countries (clade 2, *gp60* subtype IbA10G2) were
187 characterised by more stable effective population size (*Ne*=~1000). Furthermore, clade 2
188 (*gp60* subtype IbA10G2) had a higher proportion of negative Tajima's D values (Fig 2C),
189 and the distribution of Tajima's D was significantly smaller than zero (One Sample t-test: t =
190 -28.883, df = 681, p-value < 2.2e-16), which is consistent with recent population expansion.
191

192 **Linkage, recombination, introgression and gene flow between continents**

193 *Decay of linkage due to genetic exchanges*

194 To infer the rate of recombination within parasite populations from low- and high-income
195 countries, we performed independent linkage analysis for the two clades. Fig 3A indicated a
196 more rapid decay of linkage disequilibrium (LD) in clade 1 compared to clade 2 (*gp60*
197 subtype IbA10G2). This high decay of LD was consistent with genetic exchanges between
198 isolates within low-income countries resulting in genetic introgression. In contrast, the strong
199 LD in clade 2 (*gp60* subtype IbA10G2) supports our hypothesis of a recent population
200 expansion.

201

202 *Recombination and regions of secondary contact*

203 We identified four significant recombination events using the RDP4 software involving
204 genetic exchange between the two clades. These events occurred in chromosome 1 (event 1)
205 and 2 (event 2), with a further two events in chromosome 6 (event 3-4) (Fig 3B and
206 Supplementary Table 2). This indicated a potential secondary contact between geographically
207 isolated populations from clade 1 and clade 2 (*gp60* subtype IbA10G2), which resulted in

208 rare genetic exchanges between these otherwise diverged clades. Further, we detected
209 windows of selective sweeps upstream of event 3 and event 4 in chromosome 6 in clade 1.
210 These windows were further supported by regions with low π and Tajima's D (Fig 4).
211 Theoretically, this may have contributed to the recent decline in effective population size of
212 clade 1; and genetic variation would have become lost during the selective sweep as the
213 selectively favoured variant replaced other existing variants.
214 We undertook additional analyses of the highly admixed European isolate
215 (UK_UKH4 of *gp60* subtype IaA14R3) that showed unique ancestry (Fig 1B) and clustered
216 with low-income countries (clade 1). RDP4 identified four recombination events which
217 involved this admixed isolate in chromosomes 3, 4, 5 and 8 (Supplementary Fig 1 and
218 Supplementary Table 3). Moreover, a higher number of population clusters was determined
219 in these chromosomes by the Evanno method (Supplementary Fig 1), where the admixed
220 isolate had shared ancestry with these clusters (Supplementary Fig 2 & 3). The unique
221 genomic evolution of this isolate is further evident in a SplitTree graph, which showed the
222 presence of loops between clade 1 and clade 2 (*gp60* subtype IbA10G2), and in a
223 HybridCheck analysis, which revealed the signature of hybridisation which suggested
224 emergence of novel genotypes (Supplementary Fig 2 & 3).

225

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

227 To better understand the implications of genetic introgression in gene flow between the two
228 clades, we analysed recombination events in more detail. More specifically, determining the
229 signature of genetic introgression is crucial as these regions can also be responsible for
230 increasing genetic diversity and providing novel substrate for natural selection that plays an
231 important role in parasites evolution. We employed HybridCheck program [23] to perform

232 introgression analyses for recombinant events 2 – 4 (event 1 was excluded due to missing
233 parental sequence). We randomly selected a triplet (recombinant, minor parent and major
234 parental sequences) from the RDP output (Supplementary Table 2), which revealed a clear
235 signature of introgression (Fig 3 C-D), which was supported by an ABBA-BABA test (Fig
236 3E). Additionally, we calculated the pairwise r^2 between SNPs within chromosomes of clade
237 1 to determine linkage of SNPs in the introgressed regions (Fig 3F-G). Large blocks of high
238 LD that encompass the introgressed regions suggested that this region had been exchanged as
239 a single complex between clade 1 and clade 2 (*gp60* subtype IbA10G2).
240

241 **Population genetic and divergence analyses of introgressed regions**

242 We investigated whether the genetic introgression affected the genetic diversity and
243 divergence within and between clades. Hence, we calculated nucleotide diversity (π) and
244 divergence (Dxy) for each gene, within and between the two clades, respectively (Fig 5A).
245 Interestingly, we observed a pattern of high diversity and divergence particularly for genes
246 within the recombinant blocks (Fig 3B-E).

247 Next, we identified genomic islands of increased diversity and divergence and
248 investigated whether these islands were associated with genetic introgression. First, we
249 identified the 1,645 polymorphic genes between the clades (Dxy > 0) and selected the top 1%
250 genes that were highly diverged between clade 1 and clade 2 (*gp60* subtype IbA10G2)
251 (Supplementary Table 4). These top 38 genes were the most polymorphic (mean π = 0.0043)
252 compared to all remaining polymorphic genes (mean π = 0.00036) (t-test: t = -3.8055, df =
253 37, p-value = 0.0005145). Moreover, we observed that these genes were under strong
254 balancing selection (mean Tajima's D = 2.61), compared to the other polymorphic genes
255 (mean Tajima's D = 1.47) (t-test: t = -6.8032, df = 37, p-value = 4.35e-08).

256 We then explored whether the 38 divergent genes encode proteins with biological
257 properties of potential relevance for host-parasite interaction. We observed that these genes
258 were significantly enriched for extracellular encoding proteins (Chi-sq = 80.198, df = 1, p-
259 value = 0.0005), signal peptide containing proteins (Chi-sq = 10.942, df = 1, p-value =
260 0.0009) and transmembrane encoding proteins (Chi-sq = 4.6163, df = 1, p-value = 0.03167),
261 where 53%, 42%, 37% of the top 38 genes were extracellular, signal peptide and
262 transmembrane containing proteins, compared to 9%, 19%, 21% of other 1,645 polymorphic
263 genes, respectively. Secondly, 38 most divergent genes were enriched in recombinant
264 regions (Chi-sq = 287.08, df = 1, p-value = 0.00049), with 37% present in the introgressed
265 regions (Supplementary Table 4), compared to 0.9% of other 1,645 polymorphic genes.
266 Consistent with this observation, t-tests demonstrate these introgressed regions had
267 significantly elevated levels of nucleotide diversity (t = -3.0026, df = 27, p-value = 0.0057),
268 divergence (t = -3.0043, df = 27, p-value = 0.0057) and balancing selection (t = -3.0125, df =
269 27, p-value = 0.0055). Together, these results suggested that the genes undergoing frequent
270 recombination accumulated beneficial polymorphism that were maintained by balancing
271 selection, and that this increased population diversity.

272

273 *Identification of potential virulence genes*

274 Virulence genes that are involved in a coevolutionary arms race tend to be rapidly evolving.
275 Such genes are typically characterised by a relatively elevated rate of non-synonymous
276 substitutions, high intra-population variability, and/or high level of genetic divergence
277 between separate lineages. Hence, to identify potential virulence genes, we focussed on the
278 top 1% highly polymorphic genes, and we calculated the correlation between diversity (π)
279 and divergence (Dxy), and between π and Ka/Ks ratio for each gene (Fig 5B-C). We

280 identified three genes with high π, Dxy and nonsynonymous mutations encoding for mucins
281 (CHUDEA2_430, CHUDE2_440 and CHUDEA2_450) and a hypothetical protein
282 (CHUDEA6_5270) (Fig 5B-C, Supplementary Table 4). We detected episodic diversifying
283 selection (LRT p-value < 0.05) at codon potions 92, 111 and 138 in CHUDEA2_430 and at
284 codon positions 97 and 105 in CHUDEA6_5270 using the Mixed Effects Model of Evolution
285 (MEME) method implemented in Datamonkey [24]. These sites represent putative codons
286 harbouring genetic polymorphisms that experience periodicity of strong natural selection. In
287 other words, these are likely candidate virulence factors participating in a coevolutionary
288 arms race.

289

290 **Discussion**

291 In this study, we examined the evolutionary genomics of a human associated parasite, *C.*
292 *hominis*, studying whole genome sequence data from over a hundred *C. hominis* isolates from
293 16 low- and high-income countries across five continents. We thereby expand on earlier
294 studies that analysed the population genomics of *C. hominis* in Africa[19], and population
295 genetic studies that used selected genomic markers[25]. In particular, we examined the
296 evolution of two main subtype populations i.e., whole genome sequences with *gp60* subtype
297 IbA10G2, versus those that did not contain this particular *gp60* sequence ("not-IbA10G2").
298 We performed a population genomic and phylogenomic analysis to answer three questions:
299 (1) are both subtypes significantly diverged across their entire genome, (2) do they represent
300 reproductively isolated gene pools that no longer exchange genetic information through
301 recombination, and (3) do they possess genes that are significantly diverged as the result of
302 positive or balancing selection? We provide robust evidence for population diversification
303 between the two subtypes of *C. hominis* over circa 500 years of evolution. Intriguingly, we

304 revealed subtype-specific adaption and prevalence based on geographically assumed socio-
305 economic status of the infected host. We also identified genomic islands of divergence that
306 show the signatures of balancing selection. Based on our analyses, we propose to rename
307 these subtypes as subspecies to reflect their distinct evolutionary trajectory and biology.

308 Speciation is rarely discussed in human parasites despite the significance to human
309 health[26]. With improved, advanced analyses of whole genome data becoming more
310 accessible, we are now able to study the evolution and demography of parasites in much
311 detail, revealing integral differences between parasite populations that were initially
312 considered to represent a single species. Speciation plays a fundamental role in organismal
313 evolution, leading to the diversification of an organism and indeed contributes the Earth's
314 biodiversity [27]. Understanding the underlying mechanism of speciation is also equally
315 important, particularly in newly diversified species[28].

316 Phylogenetic and clustering-based analyses identified two lineages within *C. hominis*,
317 representing two spatiotemporally isolated clades. One clade (i.e., clade 1, the "non-
318 IbA10G2" subtype) comprises infections observed in low- to middle- income countries,
319 mostly from Africa and Asia, for which we propose to suffix a subspecies *C. hominis hominis*
320 referring to the fact that this subspecies is majority population. The second clade (clade 2, or
321 the IbA10G2 subtype) includes isolates from high-income countries, namely Europe, North
322 America and Oceania, and we suggest naming this subspecies *C. hominis aquapotentis*
323 (strong water), which highlights that this subspecies is mostly found in countries with
324 longstanding high sanitation and water quality indices. These two subspecies are estimated to
325 have been reproductively isolated for approximately 488 [84 – 2,199] years. The timing of
326 the divergence coincides with improvements to sanitation in Europe. Possibly, *C. h.*
327 *aquapotentis* has evolved specialisations suited for between-host transmission in higher-
328 income countries where sanitation has reduced the level of environmental transmission.

329 Indeed, epidemiological investigations in the UK identified direct person-to-person
330 transmission as a key pathway for *C. h. aquapotentis*[29]. Hence, we hypothesise this may have
331 resulted in its rapid population expansion and (partial) reproductive isolation. In contrast, *C.*
332 *h. hominis* experienced recent reduction in effective population size (*Ne*) consistent with
333 improvements to sanitation world-wide. We detected two strong signatures of a recent
334 selective sweep in this subspecies, and we propose that this may have eroded some of the
335 genetic variation, resulting in the marked drop in *Ne* = 5000 to *Ne* = 100 in the past 32 years.
336 Based on the significantly negative Tajima's D values, we infer that *C. h. aquapotentis* has
337 experienced a recent period of population expansion in the past ~500 years, providing
338 support to our hypothesis.

339 Despite increased migration and international travelling in the past decades, our PCA
340 and STRUCTURE analyses divided the global *C. hominis* isolates into three major
341 continental clusters (Fig. 1). This suggests that there is relatively little movement between
342 continents for this parasite, in contrast to that reported in *C. parvum.* Corsi *et. al.* has recently
343 showed a higher proportion of admixture and gene flow between *C. parvum* populations from
344 different geographic regions that homogenises the population structure (*in press*). Despite the
345 strong population sub-structuring in *C. hominis*, we did find evidence of potential
346 recombination between the geographic populations and subspecies. Furthermore, rapid decay
347 in LD and significant recombinant breakpoints between the populations provides evidence of
348 gene flow (Fig 3A-B). Simulation based approach had inferred 'recent geneflow' as the best
349 likelihood demographic model of *C. hominis* population, confirming secondary contact
350 between the two subspecies after recent globalisation, illustrating higher migration rate from
351 high-income to low-income countries (Fig 6). We further investigated and identified the
352 introgressed regions where we detected significant gene flow between the low- and high-
353 income countries (Fig 3C-E). This suggests that international travel can re-establish

354 connections between isolated *C. hominis* parasites, leading to secondary contact which
355 facilitates gene flow, recombination and population admixture. Potentially, this could cause
356 genetic introgression of adaptations evolved in *C. h. aquapotentis* to spread to *C. h. hominis*
357 in low-income countries, negating the progress made in sanitation in those countries with
358 respect to human health.

359 We have discovered genomic islands contributing to the population diversification,
360 and that these overlap with the introgressed regions (Fig 5A). Within the genomic islands, we
361 identified several genes that may confer a selective advantage displaying an elevated level of
362 genetic diversity under balancing selection. These genes encode proteins displaying the
363 characteristics of virulence factors being significantly enriched for extracellular, signal
364 peptide and transmembrane domain-encoding proteins and thus potentially involved in the
365 host-parasite co-evolution. Although functional data remains scarce, some data is available
366 for the mucins and *gp60* (CHUDE6_1080)[30,31]; these data suggests that by modulating
367 infectivity by sporozoites and merozoites, the level and duration of oocyst production may be
368 affected, potentially moderating virulence and resulting in the emergence of the more
369 transmissible *C. h. aquapotentis* subtypes[8,32]. In particular, we show that the orthologs of
370 mucin encoding protein CHUDE2_430 and a hypothetical protein CHUDE6_5270 are
371 consistently identified as being under positive selection (Ka/Ks > 1) in the evolution of *C.*
372 *hominis* subspecies. These findings corroborate similar conclusions made in previous studies
373 on *C. hominis* in Africa[19], and between *C. parvum parvum* zoonotic vs *C. parvum*
374 *anthroponosum* subspecies[14]. Altogether, these data strongly suggest that these genes are
375 directly related to invasion and human-host adaptation, presenting potential targets for
376 diagnostics to discriminate the subspecies.

377 In conclusion, this work represents the first large scale population genomic study in
378 any *Cryptosporidium* species, inferring the global population structure and evolutionary

379 history of *C. hominis*. We propose that the two previously identified subtypes represent
380 subspecies with their own distinct demographic histories that have diverged circa 500 years
381 ago. Although the subspecies differ in their global distribution, their gene pools are not
382 completely isolated, and rare genetic exchanges have occurred in the recent past. This
383 illustrates how human mediated gene flow is involved in parasite evolution and genomic
384 architecture, and how it could affect virulence evolution. Also, it shows that the genomic
385 islands that result from population admixture in an anthroponotic species are under selection
386 and involved in evolutionary arms race.

387

388 **Methods**

389 **Parasite isolates**

390 The *C. hominis* isolates newly sequenced for this study (n = 34) were archived stool samples
391 collected at the *Cryptosporidium* Reference Unit in the UK. The species was determined by
392 species-specific real-time PCR targeting the A135 gene[33] and the subtype by PCR and
393 sequencing of the *gp60* gene[34]. Supplementary Table 1 provides information about these
394 isolates. Isolates were selected to mainly represent the dominant variant, IbA10G2, as defined
395 by *gp60* sequencing.

396

397 **Processing of faecal samples for whole genome sequencing**

398 Stool samples were processed as previously described[35]. Briefly, saturated salt-flotation was
399 used to obtain a partially purified suspension of oocysts starting from 1-2 ml of each faecal
400 sample. Oocysts were further purified from the suspension by immunomagnetic separation
401 (IMS), using the Isolate® IMS kit (TCS Biosciences, Botolph Claydon, UK). IMS-purified
402 oocysts were treated with bleach, and washed three times with nuclease-free water by

403 centrifugation at 1,100X g for 5 min. The pellets were suspended in 200 µL of nuclease-free
404 water for DNA extraction.
405
406 **DNA preparation and whole genome sequencing**
407 Genomic DNA was extracted from purified *Cryptosporidium* oocysts by first performing eight
408 cycles of freezing in liquid nitrogen for 1 min and thawing at 95°C for 1 min, and then using
409 the QIAamp DNA extraction kit (Qiagen, Manchester, UK) according to the manufacturer's
410 instructions. The genomic DNA was eluted in 50 µL nuclease-free water, and the concentration
411 measured using the Qubit dsDNA HS Assay Kit with the Qubit 1.0 fluorometer (Invitrogen,
412 Paisley, UK), according to the manufacturer's instructions.
413 Whole genome amplification (WGA) was performed using the Repli-g Midi kit
414 (Qiagen, Milan, Italy), according to the manufacturer's instructions. Briefly, 5 µL of genomic
415 DNA (containing 1-10 ng of DNA) were mixed with 5 µL of denaturing solution and incubated
416 at room temperature for 3 min. Next, 10 µL of stop solution were added to stabilize denatured
417 DNA fragments. The reaction mixture was completed with 29 µL of buffer and 1 µL of phi29
418 polymerase, and allowed to proceed for 16 hours at 30°C. The reaction was stopped by heating
419 at 63°C for 5 minutes. WGA products were visualized by electrophoresis on a 0.7% agarose
420 gel, purified and quantified by Qubit as described above.
421 For Next Generation Sequencing (NGS) experiments, about 1 µg of purified WGA
422 product was used to generate Illumina TruSeq 2x 150 bp paired-end libraries (average insert
423 size: 500 bp), which were sequenced on an Illumina HiSeq 4000 platform (Illumina, SanDiego,
424 CA). Library preparation and NGS experiments were performed by a commercial company
425 (GATC, Germany).
426
427 **Whole genome global dataset**

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

433

434 **Data pre-processing and variant calling**

435 Raw reads of the 114 *C. hominis* isolates were trimmed to remove adapter sequences and
436 filtered for low-quality bases using Trimmomatic v.0.36[36]. The filtered reads were aligned to
437 *C. hominis* UdeA01 reference genome[37,38] using the maximal exact matches (MEM) algorithm
438 implemented in Burrows-Wheeler Alignment (BWA) tool v.0.7[39] with default settings.
439 Sequence variants were called from the aligned reads of each isolate using the HaplotypeCaller
440 method in the Genome Analysis Toolkit (GATK) v3.7.0[40] as per GATK's best practices
441 pipeline[41]. Called variants were removed if quality depth (QD) < 2.0, Fisher strand (FS) > 60.0,
442 mapping quality (MQ) < 40.0, mapping quality rank sum test (MQRankSum) < -12.5, read
443 position rank sum test (ReadPosRankSum) < -8.0, Strand odds ratio (SOR) > 4.0, QUAL < 30,
444 allele depth (AD) < 5, MAF < 0.05 or missingness > 0.5. Each of the 114 whole genomes
445 assessed here had > 80% coverage of the *C. hominis* reference genome to at least the 5-fold
446 depth. All identified variants were combined in one file and each isolate genotyped using the
447 GenotypeGVCFs tool (GATK v3.7.0) [40] and the variants were separated into SNPs (i.e.,
448 nucleotide substitutions) and INDELs.

449

450 **Population genetic structure based on whole genome SNPs**

451 The filtered SNPs were used for population structure, phylogenetic and clustering analyses.
452 Samples having multiple co-infections were identified using estMOI[42] and MOIMIX
453 (https://github.com/bahlolab/moimix). The R package SNPRelate v.1.18[43] was used for
454 principal-component analysis (PCA) analysis. A maximum likelihood phylogenetic tree was
455 constructed by IQ-TREE[44] with 1000 bootstraps and visualised in iTOL v3[45]; the sister
456 species *C. parvum* was used as an outgroup. We also constructed a consensus of 10^7 trees
457 using DensiTree 2[46] in BEAST v2[47]. BEAST v2[47] was also used to estimate the divergence
458 time between the populations by using 95% highest posterior density (HPD). A Neighbor-Net
459 algorithm based network was generated using SplitsTree5[48]. Genetic structure was analysed
460 by STRUCTURE v2.3 software[49] for population number (K) ranging 2 - 10 and plotted by
461 using plotSTR R package (https://github.com/DrewWham/Genetic-Structure-Tools). The
462 optimal population genetic cluster value K was estimated by using CLUMPAK[50].
463
464 **Population demographic history and divergence time estimation**
465 We used Bayesian Markov Chain Monte Carlo (MCMC) method implemented in Beast v2
466 program[47] to estimate the effective population size (*Ne*) of the *C. hominis* population. The
467 nucleotide substitution model of HKY was selected. A strict molecular clock model and a
468 Bayesian skyline coalescent model was used with 10^9 generations of MCMC chain and 10%
469 burn-ins. Tracer v.1.7[51] was used to assess chain convergence and effective sample size
470 [ESS] > 200 and to construct the demographic history over time; i.e. Bayesian Skyline Plot
471 (BSP).
472 Demographic histories and migration rates between the *C. hominis* populations were
473 estimated by using fastsimcoal2[52] by using mutation rate of 10^{-8} and a generation time of
474 48h[14]. We first inferred best parameters and the likelihoods for each of the demographic
475 models – no geneflow, ongoing geneflow, early geneflow, recent geneflow and different

476 geneflow, since the time of divergence (~500 years) by running 100 independent iterations
477 with 300,000 coalescent simulations and 60 optimization cycles. Demographic model with
478 the highest likelihood was then selected to run parameter estimation with block-bootstrapping
479 of 100 replicates.
480
481 **Linkage, recombination and gene flow analyses**
482 We inferred rate of decay of linkage disequilibrium (LD) by calculating the squared
483 correlation of the coefficient (r^2) between SNPs within 50kb using VCFtools[53]. LD blocks
484 were also determined by calculating pairwise r^2 between SNPs within chromosomes of each
485 population. Recombination events were identified using the Recombination Detection
486 Program version 4 (RDP4)[54] using the RDP[55], GENECONV[56], BootScan[57], MaxChi[58] and
487 Chimaera[59] methods. Events were considered significant if at least three methods predicted
488 their occurrence at a probability values, $p \leq 10^{-5}$. Recombination events with undetermined
489 parental sequences were excluded from further HybridCheck analyses. Statistically
490 significant recombination events were visualized and analyzed using HybridCheck[23] to
491 determine the sequence similarity between the isolates involved in the events. HybridCheck
492 program was also used to calculate the D statistic and estimate the gene flow between the
493 populations.
494
495 **Population genetics analyses**
496 Tajima's D, Nucleotide diversity (π), Dxy and Fst were calculated using the PopGenome R-
497 package[60]. Nonsynonymous (Ka) and synonymous (Ks) mutation rates were calculated by
498 using Ka/Ks_Calculator[61]. Protein localisation (extracellular) was predicted using WoLF
499 PSORT[62] and information regarding predicted protein targeting (signalling peptides) genes
500 were obtained from [37]. All results were plotted in R (version 3.6.1).

References

1. Khalil, I.A., *et al.* Morbidity, mortality, and long-term consequences associated with diarrhoea from Cryptosporidium infection in children younger than 5 years: a meta-analyses study. *The Lancet. Global health* **6**, e758-e768 (2018).
2. Kotloff, K.L., *et al.* Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet (London, England)* **382**, 209-222 (2013).
3. Ryan, U., Zahedi, A. & Paparini, A. Cryptosporidium in humans and animals—a one health approach to prophylaxis. *Parasite immunology* **38**, 535-547 (2016).
4. Santin, M. Cryptosporidium and Giardia in Ruminants. *Vet Clin North Am Food Anim Pract* **36**, 223-238 (2020).
5. Razakandrainibe, R., *et al.* Common occurrence of Cryptosporidium hominis in asymptomatic and symptomatic calves in France. *PLoS neglected tropical diseases* **12**, e0006355 (2018).
6. Putignani, L. & Menichella, D. Global distribution, public health and clinical impact of the protozoan pathogen cryptosporidium. *Interdisciplinary perspectives on infectious diseases* **2010**(2010).
7. Craun, G.F., Hubbs, S.A., Frost, F., Calderon, R.L. & Via, S.H. Waterborne outbreaks of cryptosporidiosis. *Journal AWWA* **90**, 81-91 (1998).
8. Feng, Y., Ryan, U.M. & Xiao, L. Genetic Diversity and Population Structure of Cryptosporidium. *Trends in parasitology* (2018).
9. Herges, G.R., *et al.* Evidence that Cryptosporidium parvum populations are panmictic and unstructured in the Upper Midwest of the United States. *Applied and environmental microbiology* **78**, 8096-8101 (2012).
10. Wang, R., *et al.* Cryptosporidium parvum IId family: clonal population and dispersal from Western Asia to other geographical regions. *Scientific reports* **4**, 4208 (2014).
11. 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).
12. King, P., *et al.* Prevalence and epidemiology of human Cryptosporidium parvum IIc infections in England and Wales. *The Lancet* **389**, S56 (2017).
13. 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).
14. Nader, J.L., *et al.* Evolutionary genomics of anthroponosis in Cryptosporidium. *Nature Microbiology* (2019).
15. 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).
16. 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).

545 17. Li, N., *et al.* Genetic recombination and Cryptosporidium hominis virulent subtype
546 IbA10G2. *Emerging infectious diseases* **19**, 1573-1582 (2013).
547 18. Gilchrist, C.A., *et al.* Genetic Diversity of Cryptosporidium hominis in a Bangladeshi
548 Community as Revealed by Whole-Genome Sequencing. *The Journal of infectious*
549 *diseases* **218**, 259-264 (2018).
550 19. Tichkule, S., *et al.* Comparative genomics revealed adaptive admixture in
551 Cryptosporidium hominis in Africa. *Microb Genom* **7**(2021).
552 20. Cacciò, S.M. & Chalmers, R.M. Human cryptosporidiosis in Europe. *Clin Microbiol*
553 *Infect* **22**, 471-480 (2016).
554 21. Feng, Y., Tiao, N., Li, N., Hlavsa, M. & Xiao, L. Multilocus sequence typing of an
555 emerging Cryptosporidium hominis subtype in the United States. *Journal of clinical*
556 *microbiology* **52**, 524-530 (2014).
557 22. Drummond, A.J., Rambaut, A., Shapiro, B. & Pybus, O.G. Bayesian coalescent inference
558 of past population dynamics from molecular sequences. *Molecular biology and*
559 *evolution* **22**, 1185-1192 (2005).
560 23. Ward, B.J. & van Oosterhout, C. HYBRIDCHECK: software for the rapid detection,
561 visualization and dating of recombinant regions in genome sequence data. *Mol Ecol*
562 *Resour* **16**, 534-539 (2016).
563 24. Pond, S.L. & Frost, S.D. Datamonkey: rapid detection of selective pressure on
564 individual sites of codon alignments. *Bioinformatics (Oxford, England)* **21**, 2531-2533
565 (2005).
566 25. Widmer, G. & Sullivan, S. Genomics and population biology of Cryptosporidium
567 species. *Parasite immunology* **34**, 61-71 (2012).
568 26. Huyse, T., Poulin, R. & Théron, A. Speciation in parasites: a population genetics
569 approach. *Trends in parasitology* **21**, 469-475 (2005).
570 27. Wolf, J.B.W., Lindell, J. & Backström, N. Speciation genetics: current status and
571 evolving approaches. *Philos Trans R Soc Lond B Biol Sci* **365**, 1717-1733 (2010).
572 28. Payseur, B.A. & Rieseberg, L.H. A genomic perspective on hybridization and speciation.
573 *Molecular ecology* **25**, 2337-2360 (2016).
574 29. McKerr, C., *et al.* Cross-Sectional Household Transmission Study of Cryptosporidium
575 Shows that C. Hominis Infections are A Key Risk Factor for Spread. (2021).
576 30. O'Connor, R.M., *et al.* Polymorphic mucin antigens CpMuc4 and CpMuc5 are integral
577 to Cryptosporidium parvum infection in vitro. *Eukaryot Cell* **8**, 461-469 (2009).
578 31. Cevallos, A.M., *et al.* Mediation of Cryptosporidium parvum infection in vitro by
579 mucin-like glycoproteins defined by a neutralizing monoclonal antibody. *Infection and*
580 *immunity* **68**, 5167-5175 (2000).
581 32. Bouzid, M., Hunter, P.R., Chalmers, R.M. & Tyler, K.M. Cryptosporidium pathogenicity
582 and virulence. *Clinical microbiology reviews* **26**, 115-134 (2013).
583 33. Robinson, G., Elwin, K. & Chalmers, R.M. Cryptosporidium Diagnostic Assays:
584 Molecular Detection. *Methods in molecular biology (Clifton, N.J.)* **2052**, 11-22 (2020).
585 34. Chalmers, R.M., Robinson, G., Elwin, K. & Elson, R. Analysis of the Cryptosporidium
586 spp. and gp60 subtypes linked to human outbreaks of cryptosporidiosis in England and
587 Wales, 2009 to 2017. *Parasites & Vectors* **12**, 95 (2019).
588 35. Hadfield, S.J., *et al.* Generation of whole genome sequences of new Cryptosporidium
589 hominis and Cryptosporidium parvum isolates directly from stool samples. *BMC*
590 *genomics* **16**, 650 (2015).

591 36. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
592 sequence data. *Bioinformatics (Oxford, England)* **30**, 2114-2120 (2014).
593 37. Heiges, M.*, et al.* CryptoDB: a Cryptosporidium bioinformatics resource update.
594 *Nucleic acids research* **34**, D419-422 (2006).
595 38. Isaza, J.P.*, et al.* Revisiting the reference genomes of human pathogenic
596 Cryptosporidium species: reannotation of C. parvum Iowa and a new C. hominis
597 reference. *Scientific reports* **5**, 16324 (2015).
598 39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
599 transform. *Bioinformatics (Oxford, England)* **25**, 1754-1760 (2009).
600 40. McKenna, A.*, et al.* The Genome Analysis Toolkit: a MapReduce framework for
601 analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303
602 (2010).
603 41. Van der Auwera, G.A.*, et al.* From FastQ data to high confidence variant calls: the
604 Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics*
605 **43**, 11.10.11-33 (2013).
606 42. Assefa, S.A.*, et al.* estMOI: estimating multiplicity of infection using parasite deep
607 sequencing data. *Bioinformatics (Oxford, England)* **30**, 1292-1294 (2014).
608 43. Zheng, X.*, et al.* A high-performance computing toolset for relatedness and principal
609 component analysis of SNP data. *Bioinformatics (Oxford, England)* **28**, 3326-3328
610 (2012).
611 44. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: A Fast and
612 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.
613 *Molecular biology and evolution* **32**, 268-274 (2014).
614 45. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and
615 annotation of phylogenetic and other trees. *Nucleic acids research* **44**, W242-245
616 (2016).
617 46. Bouckaert, R.R. & Heled, J. DensiTree 2: Seeing Trees Through the Forest. *bioRxiv*,
618 012401 (2014).
619 47. Bouckaert, R.*, et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis.
620 *PLOS Computational Biology* **10**, e1003537 (2014).
621 48. Huson, D.H. & Bryant, D. Application of phylogenetic networks in evolutionary studies.
622 *Molecular biology and evolution* **23**, 254-267 (2006).
623 49. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using
624 multilocus genotype data. *Genetics* **155**, 945-959 (2000).
625 50. Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A. & Mayrose, I. Clumpak: a
626 program for identifying clustering modes and packaging population structure
627 inferences across K. *Molecular ecology resources* **15**, 1179-1191 (2015).
628 51. Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. Posterior
629 Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic biology* **67**,
630 901-904 (2018).
631 52. Excofffier, L.*, et al.* fastsimcoal2: demographic inference under complex evolutionary
632 scenarios. *Bioinformatics (Oxford, England)* (2021).
633 53. Danecek, P.*, et al.* The variant call format and VCFtools. *Bioinformatics (Oxford,*
634 *England)* **27**, 2156-2158 (2011).
635 54. Martin, D.P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and
636 analysis of recombination patterns in virus genomes. *Virus Evolution* **1**(2015).

637 55. Martin, D. & Rybicki, E. RDP: detection of recombination amongst aligned sequences.
638 *Bioinformatics (Oxford, England)* **16**, 562-563 (2000).
639 56. Sawyer, S. GENECONV: a computer package for the statistical detection of gene
640 conversion. *http://www. math. wustl. edu/~ sawyer* (1999).
641 57. Salminen, M.O., CARR, J.K., BURKE, D.S. & McCUTCHAN, F.E. Identification of
642 breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS*
643 *research and human retroviruses* **11**, 1423-1425 (1995).
644 58. Smith, J.M. Analyzing the mosaic structure of genes. *Journal of molecular evolution* **34**,
645 126-129 (1992).
646 59. Posada, D. & Crandall, K.A. Evaluation of methods for detecting recombination from
647 DNA sequences: Computer simulations. *Proceedings of the National Academy of*
648 *Sciences* **98**, 13757 (2001).
649 60. Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E. & Lercher, M.J. PopGenome: an
650 efficient Swiss army knife for population genomic analyses in R. *Molecular biology and*
651 *evolution* **31**, 1929-1936 (2014).
652 61. Zhang, Z*., et al.* KaKs_Calculator: calculating Ka and Ks through model selection and
653 model averaging. *Genomics Proteomics Bioinformatics* **4**, 259-263 (2006).
654 62. Horton, P*., et al.* WoLF PSORT: protein localization predictor. *Nucleic acids research*
655 **35**, W585-W587 (2007).
656

657 **Acknowledgments**

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

663

664 **Author contributions**

665 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
666 the analyses. S.M.C., R.M.C., G.R., D.E., and K.M.T. were involved in acquisition of data.
667 S.T. performed the bioinformatics associated evolutionary genetic and genomic analyses.
668 A.R.J., S.M.C., C.V.O. and S.T. wrote the manuscript. All authors read and approved the
669 submission of the manuscript for the publication.

670

Orchid ID

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

Cock van Oosterhout: https://orcid.org/0000-0002-5653-738X

Simone M Cacciò: https://orcid.org/0000-0002-8561-1323

Samantha J Emery-Corbin: https://orcid.org/0000-0003-2984-0620

Kevin M. Tyler: https://orcid.org/0000-0002-0647-8158

Competing interests

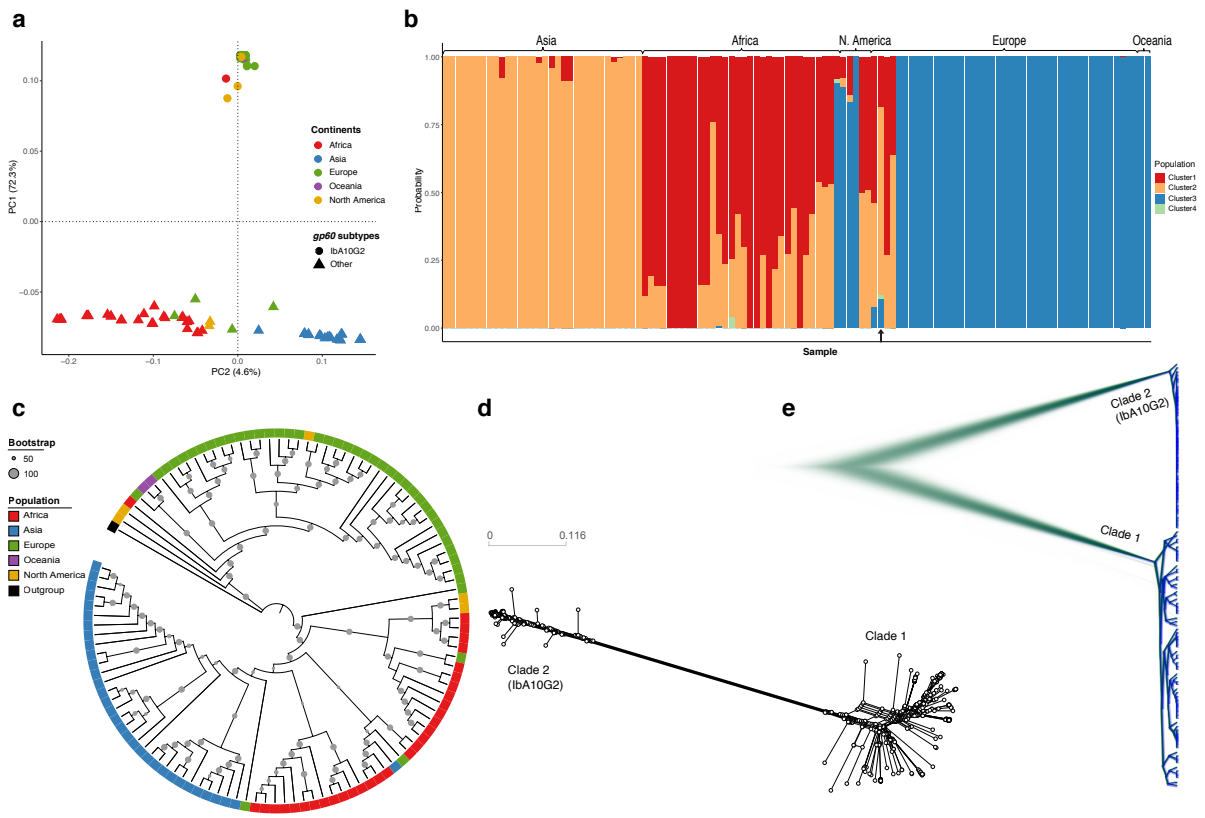
The authors declare that there are no conflicts of interest.

Funding information

A. R. J. acknowledges financial support from the Australian National Health and Medical Research Council (APP1126395) and Victorian State Government Operational Infrastructure Support and Australian Government National Health and Medical Research Council Independent Research Institute Infrastructure Support Scheme. K.M.T. was supported by funding from the Interreg 2 Seas programme 2014–2020 co-funded by the European Regional Development Fund under subsidy contract No. 2S05–043 H4DC. S.M.C acknowledges support from 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. S.T. was supported by Walter and Eliza Hall International PhD Scholarship and Melbourne Research Scholarship (MRS). C.V.O. was supported by the Earth and Life Systems Alliance (ELSA) of the Norwich Research Park (NRP).

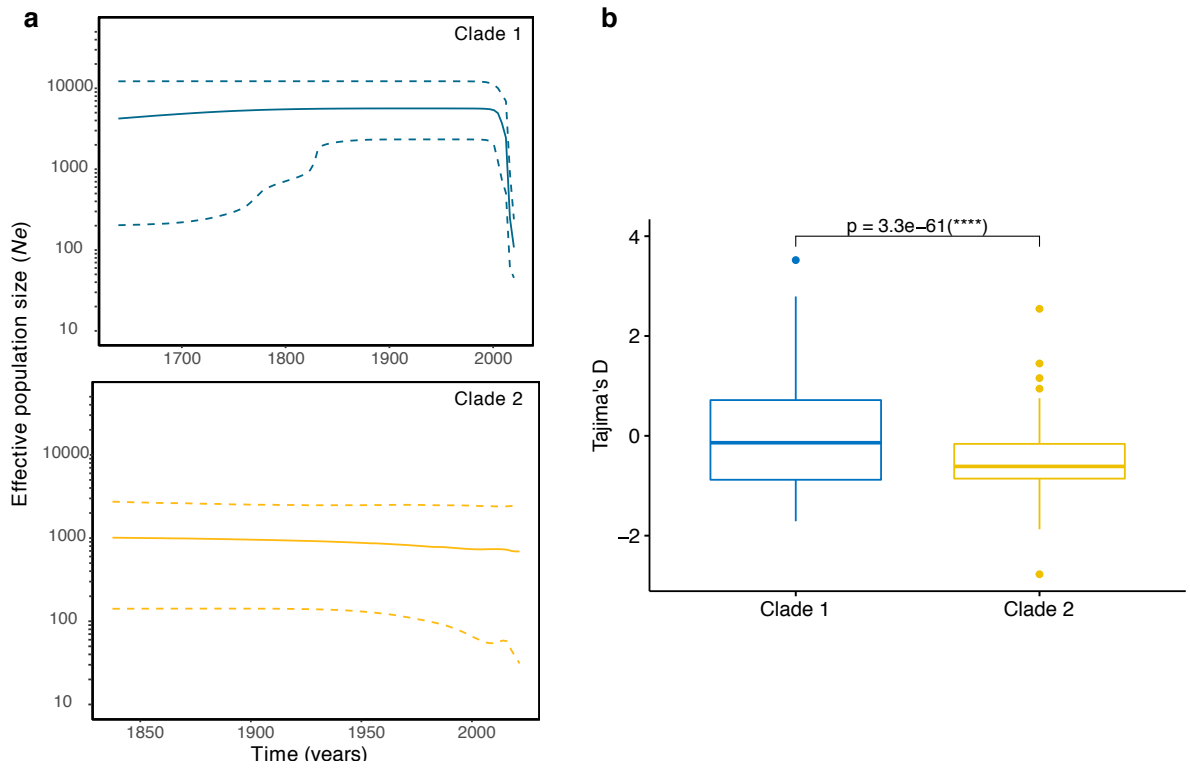
Figures

Fig 1. Global population structure of *C. hominis* isolates illustrating their sub-structing and diversification. **a**. PCA of isolates based on their 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 assuming an optimum value of K=4. The black arrow (bottom) indicates the highly admixed isolate (UK_UKH4), which includes all four ancestries. **c.** Phylogenetic clustering demonstrating two clades. **d.** Splitstree and **e.** Densitree are also demonstrating two major clades comprising isolates, which are separated based on low- and high-income countries of origin, respectively. Clade 2 includes isolates associated with *gp60* subtype IbA10G2 while clade 1 includes isolates associated with other *gp60* subtypes.



709

710 **Fig 2**. Demographic histories and population size of clade 1 and clade 2. **a**. Bayesian Skyline
711 plots (BSP) depicting change in *Ne* (effective population size) through time, for both the clades.
712 The central dark line and the upper and lower dashed lines on Y-axis are mean estimates and
713 95% HPD intervals of *Ne*, respectively. X-axis is time in years, running backwards. **b**. Boxplot
714 showing significant difference in Tajima's D values between clade 1 and clade 2.

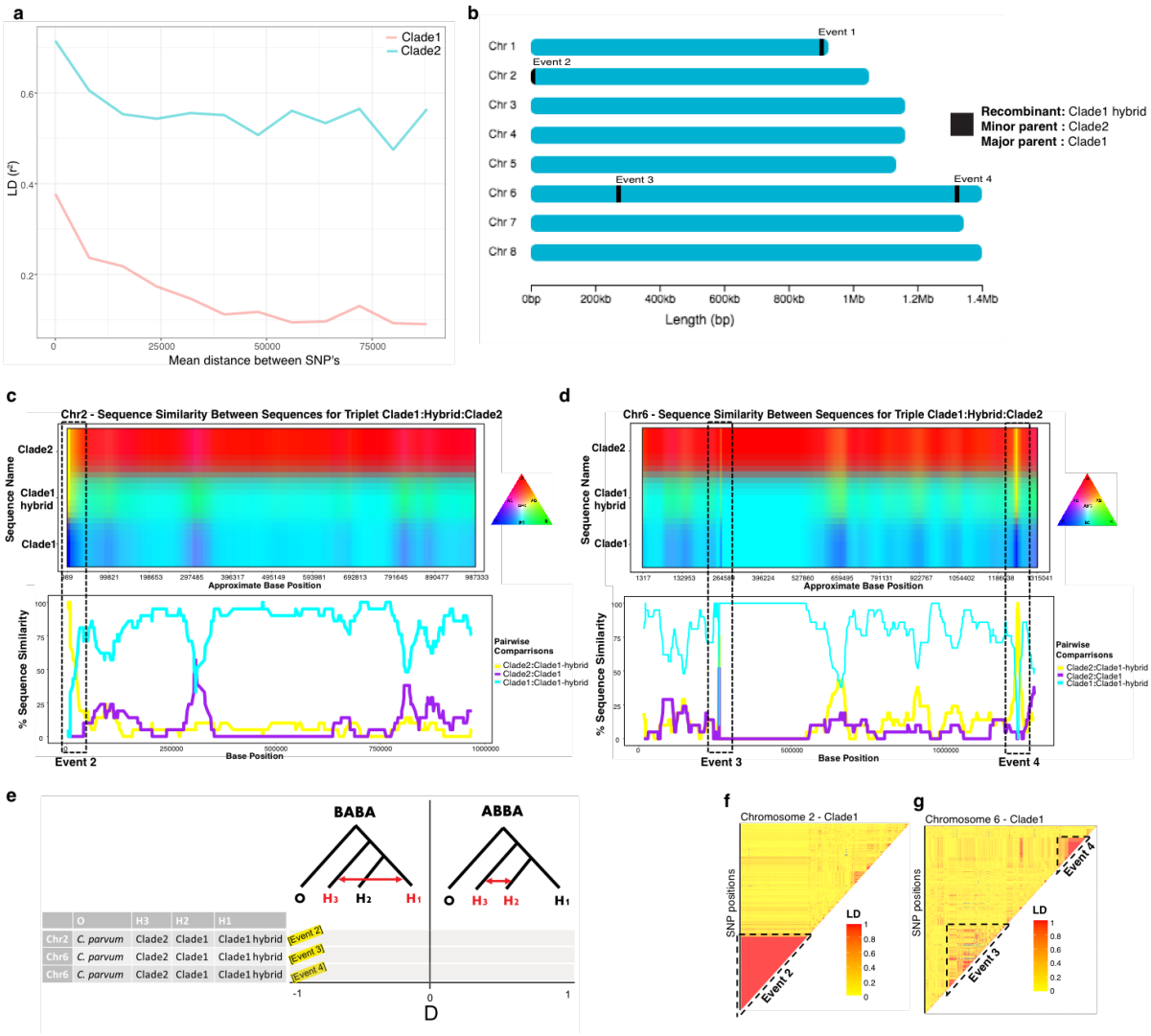


715

716

717 **Fig 3**. Analyses of recombination and gene flow between clade 1 and clade 2. **a**. Linkage
718 disequilibrium (LD) decay plot showing rapid decay of linkage between SNPs in clade 1
719 compared to clade 2. **b**. Graphical representation of recombinant breakpoint positions detected
720 by RDP4 program between clade and clade 2. **c-d**. HybridCheck plots representing genomic
721 signature of introgression in chromosome 2 and 6, respectively. Analysis for chromosome 1
722 was excluded due to unknown parental sequences. The plots were generated for random set of

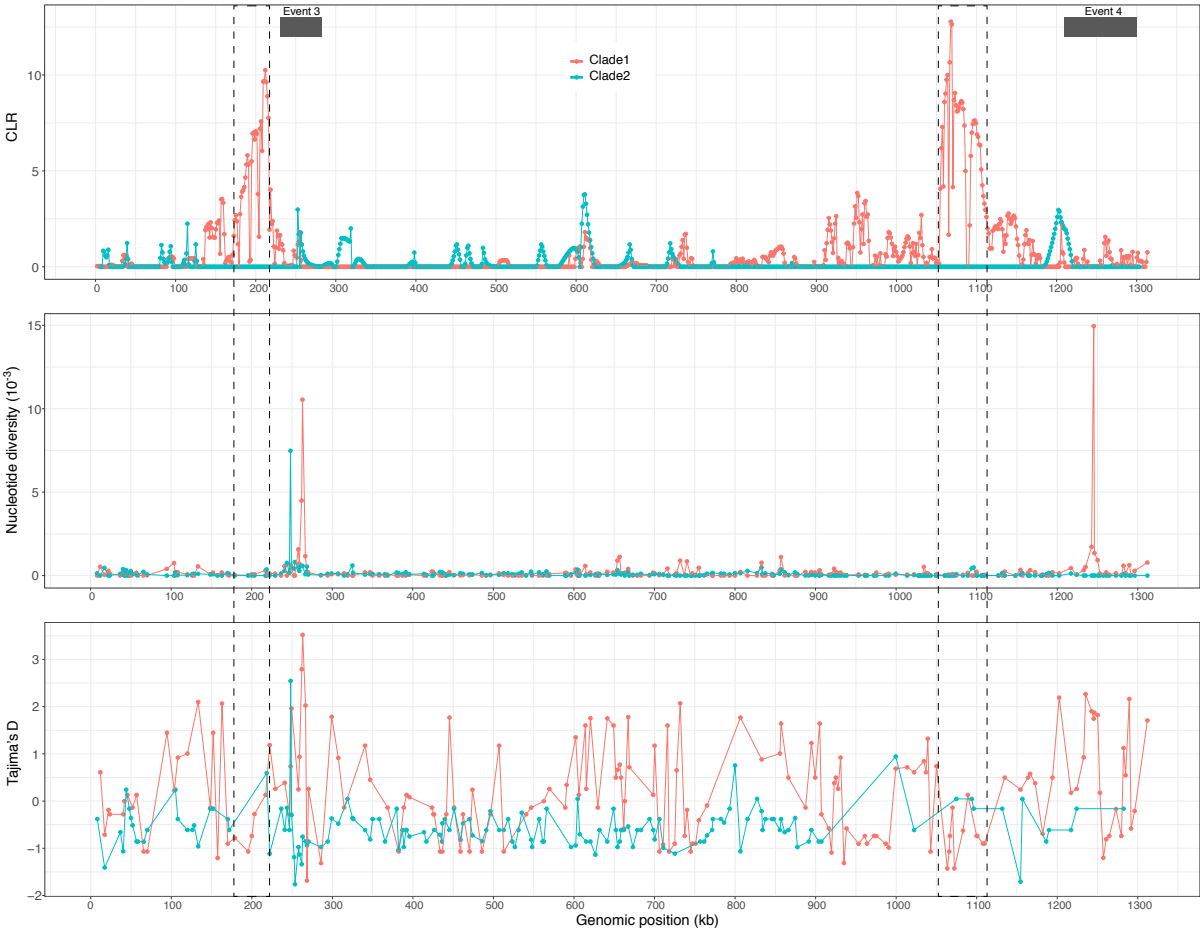
triplets that includes recombinant (hybrid), minor (donor) and major (recipient) parental sequence, as detected by RDP4 program. Introgressed blocks (recombinant breakpoints) were illustrated with dashed boxes, showing high similarity between the recombinant (clade 1 hybrid isolates) and minor parent (clade 2 isolates). **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 suggesting geneflow between H1 and H3. **f-g**. Pairwise LD of SNPs in chromosomes 2 and 6 of clade 1 showing red-blocks of high linkage between SNPs in introgressed events 2-4.



733 **Fig 4.** Signature of selective sweep in chromosome 6 of clade 1. Panels represent the CLR

734 values (top), low nucleotide diversity (middle) and negative Tajima's D values (bottom) in

735 the windows of selective sweeps.



736

737

738 **Fig 5**. Population genetic analyses of genomic islands. **a**. Population genetic and divergence

739 analyses of introgressed regions. X-axis represents genomic positions of eight chromosomes

740 highlighted with different colours. Population divergence (Dxy) between clade 1 and clade 2

741 for each gene were plotted on Y-axis (top panel). Nucleotide diversity (π) for clade 1 (middle

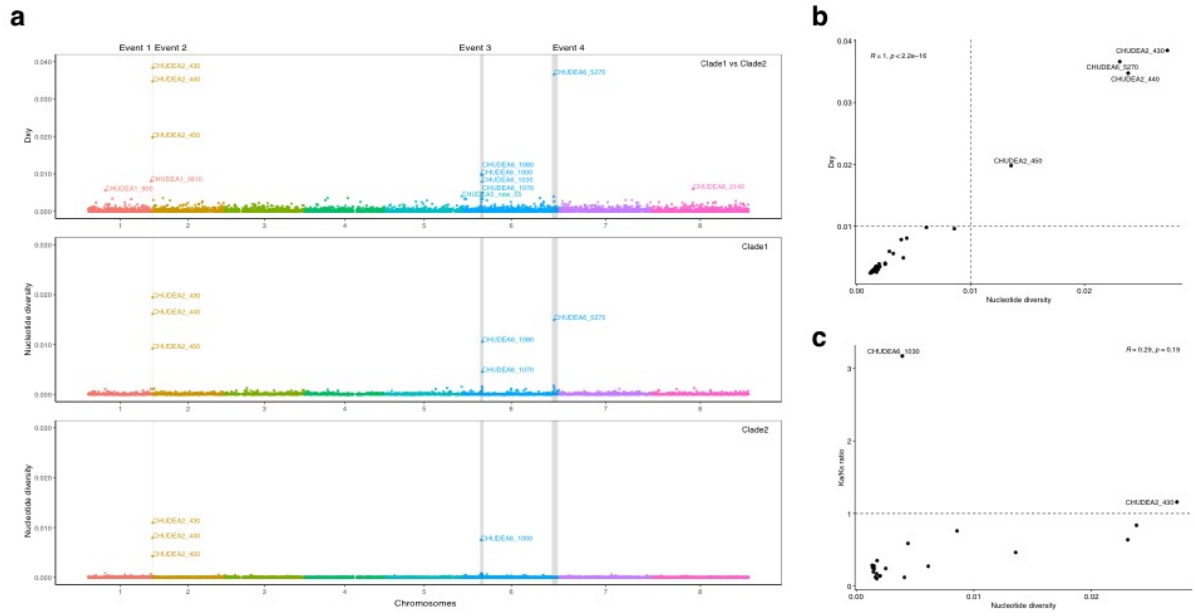
742 panel) and clade 2 (bottom panel) for each gene, were also plotted on Y-axis, respectively. The

743 breakpoints of four recombination events (event 1-4) were indicated by grey vertical boxes.

744 Event 1 was un-detected in clade 2. **b-c**. Identification of potential virulence genes. (Panel B)

745 Correlation between π and Dxy were plotted to identify polymorphic genes. (Panel C)

746 Correlation between π and Ka/Ks ratios were plotted to identify genes under selection.



747

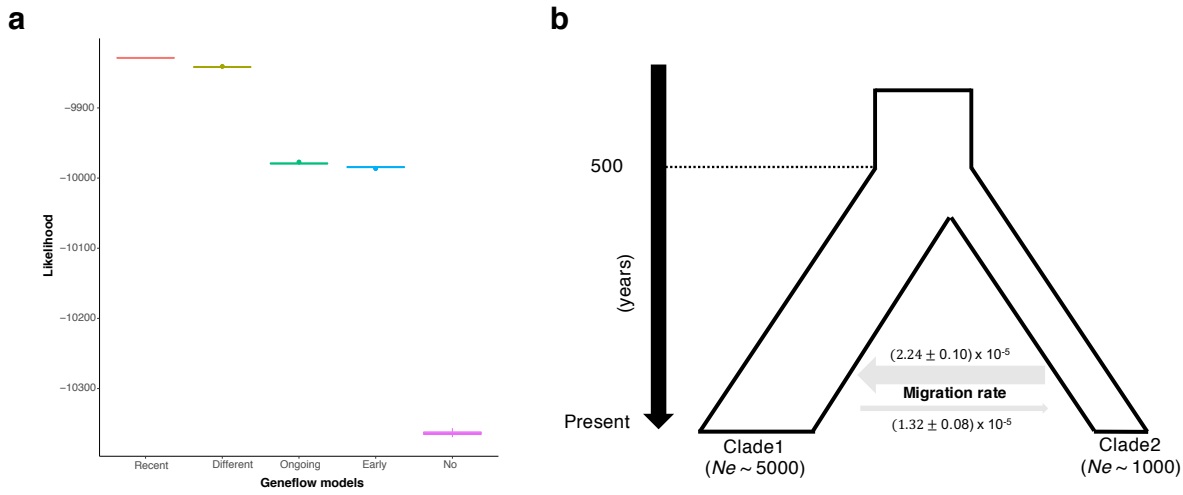
748

749 **Fig 6.** Demographic history and secondary contact between clade1 and clade 2. **a**. Comparing

750 likelihood distributions of geneflow models. "Recent geneflow" model (in red) has the highest

751 likelihood (t-test: p-value < 2.2e-16). **b**. Graphical representation of demographic history of *C.*

752 *hominis*, illustrating recent secondary contact and migration rates between the two clades.



753