GENOME-TARGETED ENRICHMENT AND SEQUENCING OF HUMAN-INFECTING *CRYPTOSPORIDIUM* spp.

Bayona-Vásquez NJ[1*], Sullivan AH[2,3], Beaudry MS[2,4], Khan A[5], Baptista RP[6,7], Petersen KN[8], Bhuiyan MIU[2], Brunelle B[4], Robinson G[9,10], Chalmers RM[9,10], Alves-Ferreira EVC[11], Grigg ME[11], Kissinger JC[3,12,13], Glenn TC[2,3,12].

[1]Division of Natural Science and Mathematics, Oxford College of Emory University, Oxford, GA, 30054, USA

[2]Department of Environmental Health Science, College of Public Health, University of Georgia, Athens, GA, 30602, USA

[3]Institute of Bioinformatics, University of Georgia, Athens, GA, 30602, USA

[4]Daicel Arbor Biosciences, Ann Arbor, MI, 48103, USA

[5]Animal Parasitic Disease Laboratory, Agricultural Research Service, US Department of Agriculture, Beltsville, MD, 20705, USA

[6]Center for Infectious Diseases, Houston Methodist Research Institute, Houston, TX, 77030, USA

[7]Department of Medicine, Weill Cornell Medical College, New York, USA

[8]Odum School of Ecology, University of Georgia, University of Georgia, Athens, GA, 30602, USA

[9]*Cryptosporidium* Reference Unit, Public Health Wales, Microbiology and Health Protection, Singleton Hospital, Swansea SA2 8QA, UK

[10]Swansea University Medical School, Singleton Park, Swansea, SA2 8PP, UK

[11]Laboratory of Parasitic Diseases, National Institutes of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, 20892, USA

[12]Department of Genetics, University of Georgia, Athens, GA, 30602, USA

[13]Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA, 30602, USA

*Corresponding author

Bayona-Vásquez and Sullivan contributed equally.

Abstract

Cryptosporidium spp. are parasites that cause severe illness in vulnerable human populations. Obtaining pure and sufficient *Cryptosporidium* DNA from clinical and environmental samples is a challenging task. Oocysts shed in available fecal samples can be limited in quantity, require purification (biased towards dominant strains), and yield limited DNA (<40 fg/oocyst). Here, we use updated genomic sequences from a broad diversity of *Cryptosporidium* species that have been found to infect humans (*C. cuniculus*, *C. hominis*, *C. meleagridis*, *C. parvum*, *C. tyzzeri*, and *C. viatorum*) to develop and validate a set of 100,000 RNA baits (CryptoCap_100k) with the aim of enriching *Cryptosporidium* DNA from varied samples. Compared to unenriched libraries, CryptoCap_100k increases the percentage of reads mapping to target genome sequences, increases the depth and breadth of genome coverage, and facilitates analyses of genetic variants in many samples, while decreasing overall costs.

46 **Introduction**

47 Cryptosporidiosis, a disease characterized by mild to severe gastrointestinal
48 symptoms, is caused by protist parasites in the genus *Cryptosporidium*. Infection can lead
49 to adverse health outcomes in vulnerable human populations (infants, the elderly,
50 immunocompromised, and pregnant)[1–6]. An estimated 7.5M global cases of
51 cryptosporidiosis occur in children aged 0–24 months, and ~133,422 deaths occur
52 annually[7–9]. In children under five years, cryptosporidiosis is the fourth leading cause of
53 diarrheal-related deaths globally[6,10,11]. In the United States, *Cryptosporidium* is the third
54 leading cause of zoonotic gastroenteritis and the leading cause of waterborne disease
55 outbreaks[12]. In low- and middle-income countries, *Cryptosporidium* are the second most
56 common pathogen in infants from 0–11 months[10].

57 There are at least 44 recognized species within the genus *Cryptosporidium*[13–15], with a
58 wide range of host specificity and likelihood of infecting humans. Thus, *Cryptosporidium*
59 spp. vary in public health importance[13]. At the genome sequence level, a cluster of
60 species shares greater than 80% identity across much of their genomes[16–19], including *C.*
61 *cuniculus*, *C. hominis*, *C. meleagridis*, *C. parvum*, *C. tyzzeri,* and *C. viatorum*.

62 Several molecular methods have been developed to gain insight into the genetic
63 diversity of *Cryptosporidium*, often relying on single-marker typing of conserved (e.g.,
64 rRNA-SSU) or polymorphic genes (e.g., *gp60* or *hsp70*)[2,20]. Other approaches include
65 multi-locus sequence typing (MLST)[21–24] and, more recently, whole-genome sequencing
66 (WGS). The disadvantages of MLST include unreliability of markers in different regions
67 due to limited information about the genetic variability of *Cryptosporidium*[24], and thus,
68 assays are not universal. Challenges with methods such as WGS are constrained by the

69 need to purify oocysts[25], given the limited amount of target DNA from fecal clinical
70 samples (which also have an abundance of other organisms' DNA) [26,27]. These challenges
71 have led to the use of purification methods combined with costly deep sequencing[28,29].
72 For example, to obtain sufficient DNA, oocysts may be propagated through animals,
73 which has been shown to change *Cryptosporidium* populations through selection of
74 individual subtypes or recombination[14,30]. Some WGS analyses have relied on whole-
75 genome amplification protocols to overcome low concentrations of target DNA[14,26,27,31].
76 A helpful alternative that still requires semi-purified oocyst suspensions[4,14,32], and that
77 may be prone to uneven amplification and the production of chimeric DNA fragments.
78 Other promising protocols for WGS directly from human stool samples do not require
79 whole-genome amplification but rely on a sufficient sample volume and an adequate
80 concentration of oocysts, achieving WGS data for ~53% of samples[33], showing medium
81 success rates. Thus, due to the limitations of existing approaches, additional methods that
82 can characterize *Cryptosporidium* genomes from limited amounts of DNA, if possible,
83 directly from fecal samples, which can be implemented easily and reliably, would be
84 beneficial.

85 Host-derived and environmental samples can also contain more than one species or
86 genotype of *Cryptosporidium* spp.[34–39]. Mixed infections may be limited to specific
87 geographic areas, particularly in endemic regions[40]. Most reported human mixed-species
88 infections contain *C. parvum* and *C. hominis*[35,37,41]; still, mixed infections are more
89 commonly represented by multiple strains of a single species[17]. Their prevalence has
90 been reported to range from 5% in the US[42] to 12% in Scotland[43], and up to 33% in
91 Peru[35]. However, their frequency and prevalence by different species or strains are

92 challenging to estimate, given the reliance on limited reference databases and PCR-
93 generated data that may fail to detect low-abundance types/species within mixed
94 infection samples[20]. Thus, methods that enable the detection and characterization of
95 mixed infections are necessary to gain a better understanding of cryptosporidiosis.

96 Hybridization sequence capture has well-known advantages[44–46] relevant to studying
97 *Cryptosporidium* spp. and other parasites: it is feasible to design thousands of baits to
98 capture genomes, such as the 9 Mb genome of *Cryptosporidium* spp.; baits can tolerate
99 mismatches (thus regions with < 20% sequence divergence may be captured); and
100 successful enrichment reduces the amount of sequencing needed (thereby reducing
101 sequencing costs in comparison with non-enriched samples). The first *Cryptosporidium*
102 whole-genome bait set (CryptoCap_75k[47]) was designed using the *C. parvum* IOWAII
103 genome assembly[48] and is reported to perform well in gene-rich regions (~90%) of the
104 genome. However, in mixed samples comprising other, more distantly related species, the
105 analyses showed a bias towards *C. parvum* enrichment over the other species[47]. Because
106 fecal samples can contain unknown or mixed *Cryptosporidium* species, a second-
107 generation bait set is needed to enrich for a broader spectrum of *Cryptosporidium* species
108 and consider highly polymorphic genomic regions to maximize the potential to capture
109 these parts of the genome.

110 To ameliorate the current challenges in the genomic study of human-infecting
111 *Cryptosporidium* spp., we developed, validated, and critically assessed CryptoCap_100k.
112 Motivated by CryptoCap_75k[47], we advance this approach by creating a bait set that also
113 captures highly diverse genomic regions (e.g., including all subtelomeric regions) and a
114 wide diversity of species that have been found to infect humans. CryptoCap_100k was

115 designed (separately from CryptoCap_75k) from the most complete publicly available
116 genome sequences from six *Cryptosporidium* species (Table 1). We validated the
117 sensitivity and specificity of CryptoCap_100k *in silico* and *in vitro* with simulated and
118 mock samples, multiple bait dilutions, and library preparation methods, including
119 enriching pools of sample libraries, rather than single libraries, for time- and cost-
120 efficiency purposes. We tested the reliability of these methods in pure oocysts and in
121 clinical samples and demonstrated that CryptoCap_100k is an efficient and economical
122 method to obtain genomic information for *Cryptosporidium*. These genomic data
123 facilitate inter-sample population genomics analyses and intra-sample genetic variation
124 via deconvolution analyses, opening new opportunities to understand the complex
125 biology of these important pathogens.
126

127 **Results**

128 *CryptoCap_100k Bait Design & Characteristics*

129 We designed a set of 100,000 120-mer RNA baits targeting six *Cryptosporidium*
130 species, including 73,672 baits for full coverage of *C. parvum* and 4,447 baits to cover
131 divergent regions among *C. hominis, C. cuniculus, and C. tyzzeri*. Additional baits were
132 designed to cover *C. meleagridis* and *C. viatorum*. Finally, we added 130 baits for genes
133 of particular interest (Table 1).

134 Mapping the baits to twelve *Cryptosporidium* genome sequences (Table S1,
135 Supplementary Notes #1) revealed ~90.5% breadth of genome coverage for target
136 species, but much lower values for distant species not considered in bait design (7.3% at
137 0.08X; Table S1; Fig. S1).

138 When comparing the multi-species design (CryptoCap_100k) to the single-species
139 design (CryptoCap_75k) (Supplementary Notes #2), despite a slightly higher percentage
140 of target reads for CryptoCap_75k than CryptoCap_100k (64.7% vs. 55.6% baits,
141 respectively), both bait sets showed similar depth and breadth of coverage (Table S1; p-
142 values > 0.05; Figs. S2–S3).
143
144 *In Silico Simulations of Enrichment and Sequencing*
145 *In silico* simulations exhibited breadth of genome coverage of up to 99.9% and a
146 depth of coverage of 7.6X (for ~200,000 paired reads), minimal off-target hybridization
147 to more than a thousand non-*Cryptosporidium* genome sequences (including the human
148 genome), and reliable mapping fidelity to the database containing ten genome sequences
149 from ten *Cryptosporidium* species (10-CrypGS), which translated to high accuracy in
150 species assignment for each simulated *Cryptosporidium* species (Supplementary Notes
151 #3–7, Supplementary Tables S2–S7, Supplementary Figures S4–S13).
152
153 *In Vitro Simulation of Single Infections and Mixed Infections*
154 DNA from oocysts of *C. parvum* and *C. meleagridis* (Supplementary Notes #8;
155 Table S8) showed improved read mapping (95.8% vs. 68.1%, p-value = 0.094; Fig. S14),
156 similar breadth (85.2% vs. 81.6%, p-value = 0.75), and greater depth of coverage in
157 enriched vs. unenriched libraries (17.3X vs. 5X, respectively; p-value = 0.01; Fig. S15–
158 S16). Species assignment was largely accurate for both library types, with a low
159 proportion of reads from *C. parvum* oocysts assigned to *C. cuniculus* and *C. hominis* (Fig.
160 S17; Table S9).

161 The proportion of reads obtained for each species (*C. parvum* and *C. meleagridis*)
162 from the unenriched mock mixed infection sample (Supplementary Notes #9; Table S10)
163 showed that initially the mixture was unequal (60.6% *C. parvum* and 39.4% *C.*
164 *meleagridis*; Fig. 1A) and enrichment slightly increased the bias (64.3% *C. parvum* and
165 35.7% *C. meleagridis*; Fig. 1A; Table S10). In total, 95.7% and 81.4% reads mapped to
166 10-CrypGS, for enriched and unenriched libraries, respectively (p-value = 0.50), with
167 very accurate species assignment for both library types (> 97%; Fig. 1B). Species
168 assignment with *gp60* performed well for *C. parvum*, but less accurately for *C.*
169 *meleagridis* (Fig. 1C; Supplementary Notes #9). Species assignments with *18S rRNA*
170 were ~70% correct (Fig. 1D; Supplementary Notes #9).
171
172 *LOD: In Vitro Sensitivity Test with Different Inputs of C. parvum DNA*
173 Dilutions of target DNA allowed the determination of the limit of detection (LOD) in
174 a background community. For NEB-iTru libraries (Supplementary Notes #10.1, Table
175 S11), the percentage of reads mapped to *C. parvum* correlated with the initial
176 concentration of target DNA, with inputs below 0.0001 ng having very low percentages,
177 with differences between 1 ng and ≤ 0.01 ng, and 0.1 ng and ≤ 0.01 ng being statistically
178 significant (p-values ≤ 0.040; Fig. S18). The average depth of genome coverage was 4X
179 for ≤ 0.01 ng, 29X for 0.1 ng, and 126X for 1 ng, with significant differences between 1
180 ng and all other dilutions (p-values < 0.0001; Fig. 2A; normalized values Fig. S19). The
181 breadth of genome coverage was significantly different between comparisons below and
182 above 0.01 ng (p-values ≤ 0.02; Fig. 2A; Supplementary Notes #10.1; normalized values
183 Fig. S19). For unenriched libraries, accurate species assignment is observed only for 1 ng

184 and 0.1 ng inputs (Fig. 2B). For enrichments, *C. parvum* is accurately detected in inputs ≥
185 0.001 ng (Fig. 2B).

186 For iNextEra libraries (Supplementary Notes #10.2; Table S12, Fig. S19), the average
187 percentage of reads mapped increased with target input in enriched libraries (Fig. S20).
188 The percentage of mapped reads was lower for inputs of ≤ 0.01 ng vs. 1 ng (p-values ≤
189 0.015) and for inputs ≤ 0.001 ng vs. 0.1 ng (p-values ≤ 0.05). Depth and breadth of
190 genome coverage tended to increase as target DNA increased (Figs. 2C and S21).
191 Although depth did not show statistically significant differences among dilutions
192 (normalized values per 100,000 reads did: Supplementary Notes #10.1, Fig. S21) for
193 libraries with ≤ 0.1 ng we recovered depths ≤ 4X whereas for those with 1 ng average
194 depth was 34.8X (Fig. 2C). Breadth of genome coverage between the two highest input
195 samples (i.e., 1 ng and 0.1 ng, mean 57%–58%) differed from all other input samples (p
196 values ≤ 0.017, Fig. 2C).

197 The average percentage of reads mapped to the *Bos taurus* genome sequence (DNA
198 used as background diluent for *C. parvum*) decreased as *C. parvum* input increased in
199 enriched libraries. With mapped reads only being significantly different for inputs of 1 ng
200 vs. all the other input values (0.004 ≤ p-values ≤ 0.021). In unenriched libraries, mapping
201 percentages to *Bos taurus* reference were high across all dilutions, averaging 90.5%
202 (Table S12, Fig. S22; Supplementary Notes #10.1).

203 Accurate species assignment was obtained in unenriched samples with inputs ≥ 0.1
204 ng, in single-enriched samples with ≥ 0.01 ng, and in double-enriched samples
205 consistently with ≥ 0.001 ng, and less consistently in samples of 0.0001 ng (Fig. 2D).
206

207 *LOD: In Vitro Sensitivity Test with Different Dilution Levels of Baits*

208 Bait dilutions enabled the determination of the minimum quantity that could be used
209 in a capture reaction. For both library types, no difference was observed among bait
210 dilutions (NEB-iTru: full down to one-sixteenth; iNextEra: one-quarter vs. one-eighth) in
211 the number of reads obtained, the percentage of reads mapped to *C. parvum,* and the
212 depth and breadth of genome coverage (Supplementary Notes #11.1–11.2; p-values ≥
213 0.05; Fig. S23–S29; Table S11).

214

215 *Enriching Seven Human Fecal DNA Samples with the CryptoCap_100k Bait Set*

216 From seven clinical patients' samples (Supplementary Notes #12.1, Table S13, Fig.
217 S30), the percentage of reads mapped increased from 0.02–0.15% (unenriched) to 0.4–
218 18.2% (enriched), with a fold change of 9.25–479X (Table S13, Fig. S31, p-value =
219 0.02). Species assignment revealed that in unenriched libraries, the few mapped reads
220 were assigned to non-human infecting *C. muris* (Fig. S32 right), similar to the
221 background effect observed in the LOD experiment (Supplemental Notes #12.2). In
222 contrast, in enriched libraries, DOCK and UKH101 are predominantly assigned to *C.*
223 *hominis* and *C. cuniculus*, while EC4 and UKP196 to *C. parvum* (Fig. S32 left). No reads
224 mapped to the *gp60* database, and very few mapped to the *18S rRNA* database with
225 assignment to species not common in human infections (Fig. S33).

226

227 *Enriching 100 Human Fecal DNA Samples with the CryptoCap_100k Bait Set*

228

229 <u>NEB-iTru libraries</u>

230 For fecal DNA obtained from 100 patients when using the NEB-iTru protocol, we
231 observed a significantly higher proportion of enriched vs. unenriched retained reads
232 mapped to the *C. parvum* genome sequence (p-value < 0.0001), increasing enrichment by
233 an average factor of 2,095X, and increasing significantly the breadth and depth of
234 coverage (p-values < 0.0008). In enriched libraries, qPCR C_T values were a good
235 predictor of the percentage of mapping reads, but not in unenriched libraries
236 (Supplementary Notes #12.2.1, Table S14; Fig. S34–S37).
237 Species assignments for unenriched libraries resemble the complex background from
238 LOD experiments (cf. Figs. 2B and S38). Following enrichment, most hits matched *C.*
239 *parvum*, but many samples have moderate proportions of reads assigned to other species,
240 including those with limited evidence to infect humans (Fig. S38; see Supplementary
241 Notes #12.2.2).
242
243 <u>iNextEra Libraries</u>
244 For the 91 clinical iNextEra libraries (Supplementary Notes #12.2.2, Tables 2 and
245 S15; Fig. S39), on average, 0.03% of unenriched reads, 15.8% of single-enriched reads,
246 and 65.0% of double-enriched reads mapped to the *C. parvum* genome sequence
247 (Supplementary Notes #12.2.2, p-values < 0.0001, Fig. 3A, Table S15, Fig. S40),
248 yielding an average enrichment factor of 1,112X from unenriched to single-enriched, and
249 an extra 9.24X from single- to double-enriched (Table S15). The mean depth of genome
250 coverage was significantly higher for double-enriched samples (95.5X) than for single-
251 enriched libraries (32.1X) and unenriched libraries (0.05X) (p-values < 0.005; Fig. 3B;
252 normalized values per 100,000 reads in Supplementary Notes #12.2.2, and Fig. S41,

253 Table S15). The mean breadth of genome coverage was significantly higher for double-
254 and single-enriched samples (48% and 54%, respectively) than for unenriched samples
255 (2.6%, p-values < 0.0001).

256 The percentage of reads mapped was negatively correlated with qPCR C_T values for
257 all library types (Figs. 3C–3E). In unenriched and double-enriched libraries, the linear
258 model fails to detect C_T values as predictors of percentage of reads mapped (p-value >
259 0.05; Figs. 3C and 3E), but in single-enriched libraries qPCR C_T has a significant but
260 small effect on percentage of reads mapping ($R^2 = 0.006$, p-value = 0.016), that is for
261 every increase of one qPCR C_T unit, the percentage decreased by 1.3% (Fig 3D). Double-
262 enrichments allowed for the retrieval of a higher percentage of target reads, even at
263 higher C_T values, than the other two library types (Figs. 3C–3E).

264 Species assignments were similar to those from NEB-iTru libraries in unenriched and
265 single-enriched libraries (cf. Figs. S38 vs. 3F and 3G). Double-enriched libraries
266 compared to single- and unenriched libraries exhibit an increased proportion of hits to the
267 most frequently observed human-infecting species (*C. parvum* and *C. hominis*; Fig. 3H).
268 Enriched libraries mapped most frequently along either *C. parvum* chromosomes for
269 samples labeled as UKP or *C. hominis*-*C. cuniculus* chromosomes for samples labeled as
270 UKH (Fig. S42), aligning with the species identified in the multi-locus typing study in the
271 same samples[49].

272

273 <u>iNextEra vs. NEB-iTru Libraries</u>

274 iNextEra and NEB-iTru libraries performed similarly (Tables 2 and S16) but with
275 noteworthy differences. In single-enriched libraries, NEB-iTru samples obtained and

276 retained more reads following quality filtering, and a higher percentage of reads mapped
277 to the *C. parvum* genome sequence than iNextEra samples. The average depth of
278 coverage is similar for both protocols (p-value = 0.7); however, the breadth of coverage
279 was significantly higher for iNextEra libraries (p-value < 0.0001; Supplementary Notes
280 #12.2.3, Table S16; Fig. S43–S47).
281
282 *Population Genomics Analyses of Clinical Samples*
283 Enrichment drastically improved SNP recovery: obtaining 94,783–127,102 SNPs for
284 double- and single-enriched vs. 15,430 for unenriched in raw VCF files (Figs. S48–S53),
285 with large variations in read depth (DP; Fig. S54Supplementary Notes #13.4). Filtered
286 VCF files contained: 76 and 77 samples, and 12,498 and 13,578 SNPs for single- and
287 double-enrichments, respectively (Figs. S55–S58). For unenriched libraries, filtering was
288 based on read depth and biallelic variants only, given that missing data filtering would
289 have removed all variants and all samples (except for three), yielding 91 samples and
290 11,226 SNPs (Figs. S59–S60; Supplementary Notes #13.4).
291 Population genomics analyses (Supplementary Notes #13.5), such as principal
292 component analysis (PCA), phylogenetic networks, genetic distance trees, and Structure,
293 are highly concordant in enriched datasets. Both datasets suggest hierarchical structure (K
294 = 2; Figs. 4, S61–S70), with distinct clustering of UKP (*C. parvum*) and UKH (*C.*
295 *hominis*) samples, aligning with the species assignment analyses from above and previous
296 multi-locus typing results[49] (cf. Figs. 3H and 4A–4B). The sparse variant database
297 obtained from unenriched libraries was insufficient for species assignment (Fig. 4C).
298 Minor distinctions between single- and double-enriched libraries are evident in sample

299 UKH149, which is positioned in the trees as more closely related to *C. parvum* than to *C.*
300 *hominis*. However, in double-enriched libraries, UKH149 clusters much closer to all
301 other *C. parvum* samples than in single enrichments (Figs. 4A and 4B).

302 High pairwise F_{ST} values between UKP and UKH samples were estimated in enriched
303 (0.94 in double vs. 0.98 in single enrichments), but not in unenriched libraries (0.08).
304 These values are balanced across all chromosomes, showing no locus bias (Figs. S71–
305 S73). Pairwise genetic distances within samples of *C. parvum* reveal two to three clusters
306 in enriched libraries (Fig. S74–S75Supplementary Notes #13.5). For *C. hominis*,
307 enrichments reveal that the three samples are highly distinct from each other, even more
308 than the differences shown between clusters of *C. parvum* (Figs. S74–S75). Genetic
309 distances using unenriched libraries fail to reflect any biologically meaningful pattern
310 (Fig. S76).

311 Deconvolution analyses (Supplementary Notes #13.6.1) revealed different numbers of
312 inferred strains (inferred K) among library types (Figs. 4A–4C). Sample UKH149 was
313 detected as having two strains by both enrichment methods (Table S17; Fig. 4B), and two
314 more samples were detected as having two strains for double-enriched libraries (UKP347
315 and UKP391) with effective Ks above 1.2 and a proportion of the least dominant
316 haplotype larger than 10% (Table S17; Fig. 4C, Supplementary Notes #13.6.1). Three
317 samples (UKH149, UKP347, and UKP391) exhibit the most differences between capture
318 methods. When assessing allelic read depth (AD) and genotype quality (GQ), UKH149
319 and UKP347 had statistically larger AD and GQ values in double-enrichments compared
320 to single-enrichments (p-values < 0.0001; Figs. S77–S80, Table S18–S19). Sample
321 UKP391 had similar AD (p-value = 0.64; Fig. S81), but GQ was statistically larger in

322 double than in single enrichments (p-values < 0.0001; Fig. S81, Table S20;
323 Supplementary Notes #13.6.2–13.6.3).
324 Genetic diversity, measured when ploidy was set to 2 in GATK (Supplementary
325 Notes #13.6.4) revealed that the nucleotide diversity (π) of polymorphic sites within
326 UKH samples (considering UKH149: single-enriched mean π = 0.066; double-enriched
327 mean π = 0.335; not considering UKH149: single-enriched mean π = 0.005; double-
328 enriched mean π = 0.0204) and within UKP samples (single-enriched mean π = 0.0122;
329 double-enriched mean π = 0.0187) was statistically different between double- and single-
330 enrichments (p-values < 0.0001); and also between UKH and UKP samples (p-values <
331 0.0002; Fig. S82). High levels of within-sample allele variation were observed in samples
332 with high values of effective K in DEploid (Supplementary Notes #13.6.5). These values
333 were consistent among chromosomes, showing no locus bias (Figs. S83–S88). Double
334 enrichments showed a higher mean of within-sample allele variation compared to single
335 enrichments (0.009 vs. 0.002), whereas unenriched libraries showed a mean within-
336 sample allele variation of 1 for all samples (Fig. S88).
337
338 **Discussion**
339 *Development and Assessment of CryptoCap_100k*
340 Obtaining sufficient *Cryptosporidium* DNA for genomic studies has been a challenge
341 in the quest to understand its global diversity and its links to transmission, infection,
342 pathogenesis, and host response. Here, we developed and evaluated a hybridization
343 capture system to obtain *Cryptosporidium* genomic data from limited amounts of input
344 DNA and evaluated its performance in a variety of settings.

345 With a foundation similar to the previous whole-genome sequencing bait capture for
346 *C. parvum* (CryptoCap_75k)[47], CryptoCap_100k used more complete genome sequences
347 that included the 16 subtelomeric regions (i.e., *C. parvum* IOWA-ATCC[50]) and included
348 additional baits to account for variation among *Cryptosporidium* species that commonly
349 infect humans. Performance between CryptoCap_75k[47] and CryptoCap_100k is similar *in*
350 *silico* (Table S1, Figs. S1–S3); however, CryptoCap_100k allows pooling ~10 libraries
351 before enrichment, reducing processing costs and time. To make recommendations about
352 how to use CryptoCap_100k best, we defined two distinct limits of detection (LOD) for:
353 i) minimum target genomic equivalents and ii) minimum bait concentration. Given that a
354 single haploid sporozoite contains ~9 fg of DNA (with four sporozoites per oocyst) and is
355 typically found in a complex community, we diluted *C. parvum* DNA in a consistent
356 background of calf DNA and determined the data quality threshold. Enriched libraries
357 exhibit a high percentage of target reads, substantial breadth and depth of genome
358 coverage, and accurate species assignment with inputs as low as 0.01 ng of target DNA
359 (~1,000 sporozoites or ~250 oocysts equivalents; Figs. 2A–2B), but higher input of target
360 DNA significantly reduces non-target reads (Fig. S22). Non-target reads are likely caused
361 by non-specific interactions of beads and/or biotin with background DNA when target
362 DNA is low[51]. These parameters improved further in double-enriched iNextEra libraries
363 with 0.001 ng of target DNA (~100 sporozoites or ~25 oocyst equivalents; Fig. 2D),
364 highlighting that double-enrichments should be prioritized for samples with very low
365 target DNA. These findings validate that CryptoCap_100k can obtain genomic
366 information at low DNA levels without purifying oocysts. The LOD assay with bait

367 dilutions suggests that there are no differences in estimated parameters between bait
368 dilutions, offering a cost-effective strategy for large-scale and effective use.
369

370 *Validating CryptoCap_100k*

371 Experiments based on simulations and on DNA from purified oocysts demonstrate
372 that CryptoCap_100k retains nearly the whole genome of the target species with minimal
373 bias (breadth of genome coverage: ~95% *in silico*, ~85% from pure oocysts; Figs. 1,
374 S15–S17). However, obtaining pure oocysts requires increased effort and time spent per
375 sample, and via propagation, strains may be altered [14,30,32]. Thus, we aimed to eliminate
376 the need for purifying oocysts to obtain usable genomic data.

377 We also successfully tested CryptoCap_100k on a small initial set of seven clinical
378 samples. Two of these had sufficient genome-wide read coverage and were also analyzed
379 with CryptoCap_75k[47], which identified UK101 as *C. hominis* and UK196 as *C.*
380 *parvum*[47], both of which agreed with our analyses (Fig. S32).

381 When validating the baits on fecal DNA from 100 clinical samples, CryptoCap_100k
382 substantially increased on-target reads relative to unenriched libraries (>437X fold
383 change) for both NEB-iTru and iNextEra protocols. While NEB-iTru libraries still
384 performed better than unenriched libraries, with a high percentage of reads mapping to
385 target genomes, iNextEra libraries achieved greater genome coverage breadth at ~50%.

386 Several studies recommend sequencing only samples with C_T values < 16[31,52],
387 however our set of 100 clinical samples had an average C_T value of 25.4. Still, we were
388 able to recover significantly more target reads, higher averages of breadth and depth of
389 genome coverage, and better accuracy of species assignment in enriched protocols (Fig.

390 S38; Fig. 3; Tables S14–S15). Double enrichments consistently resulted in significant
391 increases of target reads and depth of genome coverage (Figs. 3A–3B), as well as more
392 accurate species assignment than single-enrichments (Figs. 3F–3H). The presence of
393 some *C. cuniculus* hits in samples characterized by the multi-locus system as *C. hominis*
394 (and similar results from our *in silico* tests) underscores more likely the high genetic
395 identity between these two species[53], instead of representing mixed infections,
396 particularly given that *C. cuniculus* prevalence in the UK (where our clinical samples
397 come from) has been reported to be very low (1.2%)[54]

398

399 *CryptoCap_100k Applicability: Enables the Study of Population Genomics*

400 CryptoCap_100k is a useful tool for SNP calling, enabling robust comparative
401 analyses of genetic variation. CryptoCap_100k enabled the acquisition of genomic data
402 for downstream analyses in 76%–85% of the clinical samples used in this study.
403 Enrichment methods resolved species-level (*C. parvum* and *C. hominis*) clustering, which
404 matched the multi-locus typing results[49] (Figs. 3H, 4A–4B).

405 Beyond detecting different species, we were able to assess the diversity within each
406 species. Enrichments show that *C. parvum* haplotypes in this study are highly similar and
407 have low nucleotide diversity. Still, two clusters within the species are detected (Figs.
408 S70 and S75), which may correspond to two circulating but highly similar variants, or
409 samples with some levels of mixed infections; however, the latter was not confirmed by
410 our deconvolution analyses. Genomic studies have revealed two distinct clusters for *C.*
411 *parvum* in the UK[26]. Our samples may represent two less divergent lineages within just

412 one of these clusters, highlighting the prevalence of clonality and the dominance of some
413 lineages in the region (Figs. 4A and 4B).

414 We also identified two distinct clusters of *C. hominis* separated by the first PC axis,
415 which explains 56–57% of the variance (UKH148 + [UKH149 and UKH150]; Figs. 4A–
416 4B). Two subspecies of *C. hominis* have been detected in the UK, *C. h. hominis*, and *C.*
417 *h. aquapotentis*, with some evidence of gene flow and recombination between them in the
418 region[27]. The highly divergent *C. hominis* samples in our study may reflect the two
419 subspecies, given their comparable levels of variation. However, explicit comparative
420 analyses with samples of known lineages for both species are warranted.

421 Given that more than 90% of cryptosporidiosis cases in humans are caused by *C.*
422 *parvum* and *C. hominis*[55], our clinical results were representative. Because ~74% of the
423 baits in our set match to *C. parvum*, it is relevant to assess CryptoCap_100k performance
424 for other species. In our study, we demonstrate that CryptoCap_100k enables the
425 collection of genomic data for *C. hominis* in clinical samples, from which only ~3% of
426 the baits had been designed specifically, but we expected a larger set to work, thanks to
427 the > 97% percent similarity with *C. parvum* genome sequences[56] and the mismatch
428 tolerance of the baits. Our evidence suggests that CryptoCap_100k works appropriately
429 in other species: i) *in silico* analyses reveal proper hybridization of our baits to other
430 target species (Fig. S7); ii) *in vitro* tests using pure oocyst of *C. meleagridis* show proper
431 species assignment and a high percentage of target reads (Fig. 1B); and iii) the number of
432 baits for some species, such as *C. meleagridis* and *C. viatorum* included in
433 CryptoCap_100k is even higher than the number for *C. hominis*, which worked
434 successfully (Table 1).

435

436 *CryptoCap_100k Applicability: Potential for the Detection of Mixed Infections*

437 *Cryptosporidium* infections can contain multiple strains, subspecies, or species.
438 Current methods for detecting mixed infections include amplicon, PCR-RFLP, and
439 MLST approaches[35,57,58]. However, these methods have inherent biases due to
440 preferential amplification and can miss mixed infections that are not comprised of *C.*
441 *parvum* and *C. hominis*[35,36,57]. Regarding mixed-species infections, CryptoCap_100k was
442 able to detect these in *in silico* simulations (Fig. S13); in *in vitro* pools of *C. parvum* and
443 *C. meleagridis*, with a slight bias towards *C. parvum*; and to a lesser extent in clinical
444 samples (i.e., EC4, UKH149, UKP381, and UKH151; Figs. 3H and S32). For instance,
445 deconvolution analyses of sample UKH149 indicate a mixed-species origin, suggesting
446 the potential for the genomic data obtained via CryptoCap_100k to support the
447 deconvolution of mixed-infections within samples. Mixed infections with strains of the
448 same species were detected in UKP347 and UKP391 (Fig. 4A) in double-enriched but not
449 in single-enriched libraries. These discordances were associated with higher allelic depth
450 (AD) and/or improved genotype quality (GQ) in double-enriched libraries (Figs. S77–
451 S81), which likely allowed for greater resolution and ability to detect imbalances in
452 within-host allele frequency by DEploid. However, these mixed-species and mixed-strain
453 results should be interpreted cautiously and primarily as proof-of-concept, given the
454 limited resolution of our dataset. Also, double-enrichments exhibit higher proportions of
455 *C. parvum* compared to single-enrichments. These higher proportions of *C. parvum* can
456 be caused by minor enrichment biases due to the high number of baits designed from *C.*
457 *parvum*[47].

458

459 *CryptoCap_100k: Conclusion*

460 In summary, CryptoCap_100k results in a cost reduction of approximately 98%
461 relative to deep sequencing of fecal samples. This translates into relatively inexpensive
462 genomic data that can be applied to a large number of samples, including archived or
463 fresh samples, from diverse origins such as stool, environmental sources, oocysts, or
464 existing extracted DNA. However, cost savings will vary depending on the abundance of
465 *Cryptosporidium* DNA in the sample (Supplementary Notes #14 and Table S21).
466 CryptoCap_100k can recover >80% of the genome at high depth from pure oocyst
467 samples (Figs. S13–S14) and achieve as high as 3,690X fold-change enrichment with the
468 iNextEra protocol in fecal samples, without the need for oocyst purification, allowing
469 high-accuracy species assignment (Table S15). Also, it enables bait dilution and pooling
470 of ~10 libraries per capture reaction, which increases affordability. The quality of the
471 retrieved genomic data allows for SNPs to be called, utility for population genomics, and
472 provides the potential to assess mixed-infections. Double enrichment improves coverage
473 for low-input DNA samples, but this outcome can be matched with single enrichments
474 and higher sequencing effort. Studies using the CryptoCap_100k will enhance our
475 understanding of the population genetics of human-infecting *Cryptosporidium* species
476 and their relationship to epidemiology. Generating economical and efficient genome-
477 level data enables the growth of reference resources and the refinement of analyses and
478 methods for subtyping.

479

480 **Methods**

481 *CryptoCap_100k Bait Design & Characterization*

482 The genome sequences of six *Cryptosporidium* species were chosen for bait
483 design (Table 1), along with 55 *18S rRNA* sequences and eight *gp60* sequences (Data
484 Availability). All sequences were soft-masked for simple and low-complexity repeats
485 using RepeatMasker 4.1.1 (http://www.repeatmasker.org/). We aimed to design 100,000
486 120-nucleotide baits. Baits that were ≤ 50% soft-masked for repeats were retained. These
487 parameters resulted in 435,323 raw baits. Then, VSEARCH 2.15.0[59] was used to cluster
488 baits at various levels of sequence overlap and identity (Table 1), allowing for
489 mismatches to reduce the number of baits while maintaining diversity. The genome order
490 was designed to prioritize genome sequences based on their completeness and likelihood
491 of human infection, with *C. parvum* IOWA-ATCC listed first for all clustering
492 procedures. Then, *C. hominis*, *C. cuniculus*, and *C. tyzzeri* were collapsed at 50%
493 overlap, and 90% identity relative to *C. parvum*; and *C. meleagridis* and *C. viatorum*
494 were collapsed at 45% overlap and 85% identity relative to *C. parvum*. From *C.*
495 *viatorum*, a subset of baits was randomly selected. Raw baits were obtained for the *18S*
496 *rRNA* targets and collapsed at 50% overlap and 95% identity (Table 1).

497 We characterized CryptoCap_100k (Supplementary Notes #1) by using twelve
498 genome sequences available for ten species of *Cryptosporidium* (Data Availability).
499 BWA 0.7.17[60] was used to map the CryptoCap_100k bait sequences to each genome
500 sequence. Alignment files were generated and converted using Samtools 1.10[60,61]. The
501 bamCoverage function from DeepTools 3.3.1[62] was used to produce bigwig files to
502 visualize coverage tracks calculated as the number of reads per bin in IGV-Web app

503 version 1.10.8[63]. To obtain alignment statistics, the depth option from Samtools v1.1.0
504 and the *-a* flag were used to get all sites and create a depth file[61]. The mean depth, the
505 proportion of baits mapped to each reference, and the breadth of genome coverage were
506 estimated. The breadth of coverage, as reported throughout this manuscript, was
507 determined by estimating the number of bases covered in the alignment over the
508 reference length, with a minimum coverage depth of 1X.
509 We also compared the CryptoCap_100k bait set to CryptoCap_75k[47]
510 (Supplementary Notes #2) by estimating the same parameters as above, and using
511 Welch's two-sample t-test for comparisons.
512
513 *In Silico Simulations of Enrichment with CryptoCap_100k & Sequencing*
514 To validate data quality from CryptoCap_100k before bait synthesis, a simple
515 infection scenario, error-free and bias-free, was simulated. We used twelve available
516 genome sequences from ten *Cryptosporidium* species to simulate the enrichment and
517 sequencing process with CryptoCap_100k, assuming two different ranges of insert size
518 fragments (Supplementary Notes #3). To simulate hybridization, we used the coordinates
519 of the mapped baits in relation to each reference genome, and we extended the position
520 by several bases to "capture" larger fragments than the bait length using Samtools[61]. To
521 simulate sequencing, ART 2016.06.05[64] was used to reproduce ~200,000 paired-end 150
522 bp fastq reads from the extended coordinates in each reference genome. Then, reads were
523 mapped to their respective genome sequences using BWA 0.7.17[60], and we estimated
524 depth and breadth of coverage. To validate the accuracy of species assignment, we used a
525 reference database containing ten *Cryptosporidium* species genome sequences (10-

526 CrypGS), a *gp60* database[2], and an *18S rRNA* database[55,65,66], and used BBMap 38.90[67] to
527 map simulated fastq reads to each and explore the proportion of hits to each
528 *Cryptosporidium* species.

529 To assess the sensitivity of the baits to off-target DNA, we assessed the affinity of
530 our baits to 1,008 genome sequences[68], including the human genome. Because *in vitro*
531 libraries were diluted in calf thymus DNA (see *In Vitro Test of Limits of Detection with*
532 *Different Levels of C. parvum Dilution*), we also assessed affinity to the *Bos taurus*
533 genome sequence (GCA_002263795.3). For both analyses, we used BWA 0.7.17[60] to
534 map the bait sequences, and we estimated the number of hits as well as the quality of the
535 mapping.

536 To explore the ability to discriminate different species within mixed-species
537 infection samples, we created four simulated samples by concatenating the fastq files
538 already simulated for each species above. The four *Cryptosporidium* mixed infection
539 samples contained at equal proportions: i) reads from twelve genome sequences from ten
540 species; ii) reads from ten genome sequences from ten species; iii) reads from five
541 genome sequences and species that share high percent identity; and iv) reads from three
542 genome sequences and species commonly found in human infections. Then, we mapped
543 these mixed reads to 10-CrypGS, the *gp60*, and the *18S rRNA* databases. See
544 Supplementary Notes #3–7 for details.

545

546 *In Vitro Enrichment & Validation with CryptoCap_100k Baits with Pure Oocysts' DNA*

547 Pure oocyst DNA was obtained from three *C. parvum* samples from Dr. Michael
548 Grigg (National Institutes of Health, US) and *C. meleagridis* DNA isolate TU1867 from

549 BEI Resources (BEI Resources NR2521 Manassa, VA). Metagenomic libraries were
550 prepared with the New England BioLabs Ultra FS II library kit (see *Metagenomic*
551 *Libraries*), and aliquots of these were left unenriched or enriched with one-tenth of the
552 recommended volume of CryptoCap_100k (see *Hybridization Capture Enrichments*).
553 Libraries were sequenced using an Illumina MiSeq with a PE250 kit.
554
555 *In Vitro Simulation of Mixed Infections*
556 To test the ability to detect mixed infection samples, after NEB/iTru library
557 preparation (see *In Vitro Enrichment & Validation with CryptoCap_100k Baits with Pure*
558 *Oocysts' DNA*) but before bait capture, two indexed libraries were pooled at equimolar
559 ratios to simulate a mixed infection. After pooling, an aliquot was enriched with one-
560 tenth of the recommended volume of CryptoCap_100k, and another remained
561 unenriched. These were sequenced in an Illumina MiSeq using a PE250 kit. We
562 estimated the number of reads per sample after demultiplexing, expecting roughly 50%
563 for each species.
564
565 *In Vitro Test of Limits of Detection with Different Levels of C. parvum Dilution*
566 To determine the CryptoCap_100k limits of detection (LOD) of *Cryptosporidium*,
567 we performed two types of library preparations, NEB-iTru and iNextEra[69], on samples
568 with known and varied input DNA. For NEB-iTru libraries, we diluted *C. parvum*
569 genomic DNA in a background of calf thymus DNA (Sigma-Aldrich, St. Louis, MO). A
570 45-ng aliquot of calf thymus DNA was pipetted into five tubes. In the first tube, 5 ng of
571 *C. parvum* genomic DNA was added to a final volume of 50 µL. Then, three 1:10 serial
572 dilutions were performed. The fifth tube corresponded to a negative control with just the

573 calf DNA. Metagenomic libraries using the NEB-iTru protocol and 10 µL of each
574 dilution were generated (see *Metagenomic Libraries*). In this experiment, up to 7 libraries
575 were pooled, an aliquot was left unenriched, and independent aliquots were enriched with
576 full, one-half, a quarter, an eighth, or a sixteenth of the recommended volume of
577 CryptoCap_100k (see *Hybridization Capture Enrichments*). Pools were sequenced at
578 Novogene using a HiSeq X with a PE150 kit. The number of reads, percentage of reads
579 mapped to the *C. parvum* genome sequence, and depth and breadth of genome coverage
580 were estimated.

581 For iNextEra libraries, 90 ng aliquots of calf DNA were pipetted into six tubes. In
582 the first tube, 10 ng of *C. parvum* genomic DNA was added in a final volume of 100 µL.
583 From this, four 1:10 serial dilutions were performed. The sixth tube corresponded to the
584 negative control (calf DNA only). iNextEra library preparation was performed with 10
585 µL inputs from these dilutions (see *Metagenomic Libraries*). Up to 6 libraries were
586 pooled, and aliquots were left unenriched or enriched with one-quarter or one-eighth of
587 the recommended baits. Aliquots of single-enriched libraries were double-enriched with
588 one-quarter of the recommended baits (see *Hybridization Capture Enrichments*). Pools of
589 unenriched, single-enriched, and double-enriched libraries were sequenced at Novogene
590 using a NovaSeq with a PE150 kit. The number of reads, the percentage of reads mapped
591 to the *C. parvum* genome sequence, the depth and breadth of genome coverage, and
592 species assignment based on the mapping of reads against 10-CrypGS were analyzed
593 among dilutions.

594

595 *In Vitro Test of Limits of Detection with Different Levels of Bait Dilution*

596 To determine the sensitivity or limits of detection (LOD) when using a lower
597 concentration of baits than recommended, we modulated the quantity of baits added to
598 the capture reaction in one experiment with NEB-iTru libraries and one experiment with
599 iNextEra libraries.

600 For both library types, the bait dilution sensitivity assay considered pools of the
601 metagenomic libraries from the *In Vitro Test of Limits of Detection with Different Levels*
602 *of C. parvum Dilution* experiment. For NEB-iTru libraries, we tested five different bait
603 quantities: full volume (5.5 µL, per the manufacturer's specifications), and dilutions in
604 water at one-half, a quarter, an eighth, and a sixteenth. For iNextEra libraries, we tested a
605 quarter and an eighth bait dilutions. Sequencing was performed as described in the
606 section above. The number of reads, the percentage of reads mapped to the *C. parvum*
607 genome sequence, and the depth and breadth of genome coverage were compared among
608 bait dilutions.

609

610 *In Vitro Enrichment & Validation with CryptoCap_100k Baits with Clinical Samples*

611 Two clinical datasets were used for real-world validation. First, Dr. Michael
612 Grigg (National Institutes of Health, US) and Prof. Rachel Chalmers's laboratory (Public
613 Health Wales—PHW— Microbiology and Health Protection, UK) provided DNA
614 extracted (Qiagen PowerSoil Kit) from seven de-identified patient fecal samples. These
615 samples were obtained under an NIAID IRB-approved protocol NCT00006150. The
616 NEB-iTru library protocol was used, pools made, aliquots were either left unenriched or
617 enriched with one-tenth of the recommended volume for baits, and libraries were
618 sequenced in an Illumina HiSeq using a PE150 kit.

619 Second, 100 DNA extractions (Qiagen QIAamp Fast DNA Stool kit) from clinical
620 samples from Prof. Rachel Chalmers's laboratory (PHW) were used. This work was
621 carried out in accordance with a material transfer agreement between UGA and PHW,
622 which did not require ethical approval. These samples had previously been screened
623 using a multi-locus typing system [49] to detect the species present, and these previous
624 results are reflected in the sample names used here (*C. parvum* = UKP and *C. hominis* =
625 UKH). The laboratory workflow used for these clinical samples in this study is depicted
626 in Fig. 5. A real-time qPCR assay was used to quantify *Cryptosporidium*, using primers
627 for the *18S rRNA* gene and slightly modified probes [70] as specified in Supplementary
628 Notes #12.2. Two types of genomic libraries were prepared for each sample (NEB-iTru
629 and iNextEra; see *Metagenomic Libraries*). Nine DNA samples were depleted when used
630 for iTru libraries, thus not used for iNextEra libraries. Aliquots of pools of ~12 libraries
631 were left unenriched or enriched. iTru-NEB libraries were enriched with one-tenth and
632 iNextEra libraries were enriched with one-eighth of the recommended volume of
633 CryptoCap_100k. For iNextEra libraries, we performed additional double enrichments on
634 single-enriched aliquots using one-quarter of the recommended volume of baits (see
635 *Hybridization Capture Enrichments with CryptoCap_100k*). All libraries were sequenced
636 using an Illumina NovaSeq X with a PE150 kit (Fig. 5).
637
638 *Metagenomic Libraries*
639 For iTru-NEB libraries, the New England BioLabs (NEB) Ultra FS II library kit
640 (New England BioLabs, Ipswich, MA) protocol was used according to the manufacturer,
641 with enzymatic fragmentation lasting 5–11 minutes, based on the integrity of DNA in gel

642 electrophoresis. A 12- or 14-cycle PCR with iTru primers[69] was used to add unique dual-
643 indexes to each ligation product. Products were cleaned with 1.25X SpeedBeads[71] and
644 quantified with a Qubit 4.0 Fluorometer (Thermofisher, Waltham, MA).

645 iNextEra library preparation was used to decrease the time, cost, and amount of
646 input DNA needed. In brief, 2.5 µL of 10 ng/µL (25 ng) of sample DNA was used for a
647 tagmentation reaction with 3 µL of 2X TMP buffer and 0.5 µL of Transposome (BLT;
648 CAT # 20015880, Illumina Inc.), incubated at 53 °C for 30 minutes. Then, PCR reactions
649 contained 9.5 µL of Q5 Buffer, 1.25 µL of 10 mM dNTPs, 0.5 µL of Q5 High Fidelity
650 Polymerase (New England Biolabs M0491L), 40.25 µL of nuclease-free water, 2.5 µL of
651 iNextEra5 indexed primer at 5 µM, 2.5 µL of iNextEra7 indexed primer at 5 µM[69], and 6
652 µL of tagmented product. Cycling conditions were: 3 minutes at 72 °C, 3 minutes at 98
653 °C, followed by ten cycles of 98 °C for 45 seconds, 62 °C for 30 seconds, and 72 °C for 2
654 minutes, with a final extension at 72 °C for 1 minute.

655

656 *Hybridization Capture Enrichments with CryptoCap_100k*

657 Metagenomic libraries were pooled (e.g., Fig. 5) in preparation for enrichment.
658 For different experiments, the input concentration per pool of samples varied (i.e., for
659 clinical samples using NEB/ iTru, the input was between 25–80 ng/µL; for clinical
660 samples using iNextEra, it was 103 ng/µL; and for the LOD experiments, it was 26–59
661 ng/µL). The enrichment process per pool followed manufacturers' standard protocols
662 (mybaits v5, Daicel Arbor Biosciences, Ann Arbor, MI), with wash and hybridization
663 temperatures at 63 °C and modifications of bait quantities as described above. After

664 enrichments, P5/P7 PCRs with 16–22 cycles were performed, followed by 1–1.25X
665 SpeedBeads clean-ups (Sigma Aldrich, St. Louis, MO).
666 For the 100-clinical sample set and the LOD experiments, double enrichments
667 were performed on the iNextEra single-captured pools (e.g., Fig. 5). For these, aliquots of
668 the cleaned single-captured pools were pooled and used independently as input for a
669 second capture using one-quarter of the recommended baits. For post-double-enrichments
670 P5/P7 PCRs with 12–16 cycles were performed, followed by 1X SpeedBeads clean-up.
671 All PCRs were performed using the KAPA HiFi Kit (Roche, Basel, Switzerland). All
672 P5/P7 PCRs consisted of 98°C for 2 minutes.; then, cycles of: 98°C for 20 sec., 60°C for
673 30 sec., 72°C for 1 min.; followed by 72°C for 5 min.
674
675 *In Vitro Bioinformatic Processing and Analyses*
676 Bioinformatic and data analysis code are available in the Supplementary Notes.
677 Sequencing data were processed using Trimmomatic 0.39[72] , and retained reads were
678 mapped to specific genome sequences or databases using BBMap 38.9[67]. Samtools[61] was
679 used to sort bam files. We determined mapping statistics, including genome depth and
680 breadth of coverage. The results of the mapping to 10-CrypGS were used to assess the
681 proportion of hits to the different *Cryptosporidium* species and produce species
682 assignment plots. For iNextEra libraries from clinical samples, we also generated bigwig
683 files, which were visualized using the IGV-Web app version 2.2.7.
684 Cleaned reads from pure oocysts libraries and *in vitro* mixed-infection libraries
685 were mapped to the corresponding target references (i.e., *C. parvum* samples to *C.*
686 *parvum* IOWA-ATCC, *C. meleagridis* samples to chromosome-level assembly[73], also to

687 10-CrypGS, to the *gp60* database, and the *18S rRNA* database—used in simulations). The
688 sequencing reads from the LOD dilution experiments (i.e., input DNA and bait dilutions)
689 were mapped to *C. parvum* IOWA-ATCC and 10-CrypGS. iNextEra reads from LOD
690 were also mapped to *Bos taurus* (GCA_002263795.3). The cleaned sequence reads from
691 clinical samples were mapped to *C. parvum* IOWA-ATCC and 10-CrypGS. For the set of
692 100 clinical samples, qPCR C_T values were used in the *parsnip 1.1.1*[74] to define linear
693 regression models predicting the percentage of reads mapped against 10-CrypGS
694 (Supplementary Notes #12).
695
696 *Population Genomics and Mixed Infection Analyses of Clinical Samples*
697 To examine the genetic variation in natural populations of *Cryptosporidium* from the
698 set of 91 clinical samples prepared using the iNextEra protocol, we performed genotype
699 calling (Fig. S89). In short, reads from each treatment (unenriched, single-enriched, and
700 double-enriched) were mapped to the *C. parvum* IOWA-ATCC genome sequence using
701 BWA 0.7.17[60] , and the results were used in GATK 4.6.0.0[75]. Duplicate reads and
702 haplotypes were determined using Haplotypecaller with GVCF mode and merged into
703 databases for each library type to perform joint genotyping using GenotypeGVCFs.
704 Importantly, HaplotypeCaller was run with --sample-ploidy set to 1. A first filtering
705 approach was performed to obtain SNPs with QD < 2, FS > 60, MQ <40, MQRankSum <
706 -12.5, and ReadPosRankSum < -8.0.
707 The resulting VCFs were explored for missing data and coverage depth per sample
708 and chromosome using vcfR 1.15.5[76]. Then, filtering using dartR 2.9.7[77] considered only
709 SNPs with coverage (DP) larger than 5 and lower than 200, that were biallelic in the

710 dataset, and with less than 40% of missing data. Individuals with > 80% missing data
711 were removed.
712 To explore the diversity and divergence among samples, filtered VCF files for each
713 treatment were used for population genomic analyses. These included Principal
714 Components Analyses performed with SNPRelate 1.40.0[78] and ade4 1.7-23[79]. We used
715 phangorn 2.12.1[80]and SplitsTree 6.4.11[81] to compute pairwise genetic distances and
716 visualize phylogenetic networks. We estimated Nei's genetic distances between samples
717 and visualized them in dendrograms with bootstrap support using poppr 2.9.6[82] and
718 plotted these with ape 5.8-1[83]. We tested K = 1–5 clusters, each with three replicates, in
719 Structure 2.3.4[84], run through structure_threader 1.3.10[85], which utilizes Structure
720 Harvester 0.6.94[86]. We used 10,000 chains for burn-in, 100,000 chains for MCMC, and
721 set ploidy to 1, with other options left as default. We used readr 2.1.5[87] and starmie
722 0.1.2[88] to graph Evanno's results and to plot Structure assignment plots. We estimated
723 pairwise F_{ST} statistics between UKP and UKH samples using hierfstat 0.5-11[89] and
724 mapped these values along *C. parvum* chromosomes using qqman 0.1.9[90]. We compared
725 Allelic depth (AD) and genotype quality (GQ) values between VCF files of the different
726 treatments and used ggridges 0.5.6[91] to plot the frequency of values across samples.[90]
727 To explore the diversity within each clinical sample, we deconvoluted multiple
728 infecting strains using DEploid 0.7.1[92]with no reference panel. The effective and inferred
729 number of strains (K) was plotted. Additionally, using filtered VCFs from the GATK
730 4.6.0.0[75] run with the setting --sample-ploidy set to 2, we estimated nucleotide diversity
731 (π) of SNPs in VCFTools 0.1.16[93] for double and single enrichments that were subset to
732 contain only UKH (*C. hominis*) or UKP (*C. parvum*) samples; and, we assessed and

733 plotted within-sample allele variation (what in a diploid system would be categorized as
734 heterozygosity) levels per sample. Data analyses were performed in R Studio 2023.06.0.
735 Least-square comparisons of the means between experimental groups (e.g., unenriched
736 and enriched, single-capture vs. double-capture, etc.) were performed to calculate p-
737 values using a confidence interval of 0.95 with the package: *lsmeans*[94], which provides a
738 t-test output when comparing two variables, or performs a Tukey method adjustment
739 when comparing a family of three or more estimates. Other packages used were *tidyverse*
740 *2.0.0*[95], *ggplot2 3.5.2*[96], and *plotly 4.10.4*[97].
741

742 **Acknowledgments**

743 T.G. and J.C.K. disclose support from the National Institutes of Health under
744 award number NIH R01AI148667. M.G. discloses support from the Intramural Research
745 Program of the National Institute of Allergy and Infectious Diseases (NIAID) at the
746 National Institutes of Health. We thank the authors of CryptoCap_75k[47] for sharing
747 unpublished results that facilitated the work and comparisons presented here. We also
748 thank Prisha Sharma and Fiifi Agyabeng-Dadzie for sample processing, Christopher
749 Fitzgerald for sample tracking and processing, and Mustafa Nural for sample database
750 help. This study was supported in part by resources and technical expertise from the
751 Georgia Advanced Computing Resource Center, a partnership between the University of
752 Georgia's Office of the Vice President for Research and Office of the Vice President for
753 Information Technology.
754

755 **Author Contributions**

N.J.B-V., A.H.S., M.S.B., A.K., R.P.B., M.E.G., J.C.K., and T.C.G. contributed to conceptualization, methodology, investigation, and writing—review and editing. M.S.B., R.P.B., and B.B. designed and oversaw the production and QC of the baits. M.S.B., K.N.P., M.I.U.B., contributed to investigation and writing—review and editing. N.J.B-V. and A.H.S. performed validation, data curation, formal analyses, and with T.C.G. writing—original draft. G.R., R.M.C., E. A-F, M.E.G., and J.C.K. contributed samples and analyses to the manuscript. N.J.B-V, M.S.B, A.K., R.P.B., M.E.G., J.C.K., and T.C.G contributed to funding acquisition. All authors read and approved the final version of the manuscript.

Competing Interests

B.B. and M.B. are employed by Diacel Arbor Biosciences, which manufactures and sells the baits described herein. The CryptoCap_100k bait set is available for commercial purchase (Design ID: D10006Crypto, Cat# #308508.V5, #308548.V5, and #308596.V5, Arbor Biosciences, MI, USA).

Data Availability

The six genome sequences that supported the CryptoCap_100k bait design were from GenBank: GCA_015245375.1, GCA_004337795.1, GCA_001483515.1, GCA_007210665.1, GCA_004337835.1, and GCA_001593445.1. The extra sub-telomeric sequences from *C. parvum* were: MZ892386.1, MZ892388.1, and MZ892387.1.

778 The *18S rRNA* sequence data that supported the CryptoCap_100k bait design
779 were from GenBank with the EU250845, LC089976, DQ288166, EF641022, GU951714,
780 AF108864, LC012016, GQ121020, HQ397716, KR296812, EF641014, FJ435960,
781 JX416368, KP099082, AY120903, AY120902, JX644908, AY324641, AY324638,
782 AF159113.1, JQ413438, AF108861, GQ121021, KP730304, AY737602, AY731235,
783 JX294358, AY268584, KR819168, AY120904.1, KU608308, AF112576, EU553551,
784 JF710247, KP004205, JQ002555, KT235702, AB909498.1, AF316630, AY504515,
785 AY504516, AF262325, GQ227479, EU827297, AF093500, AF093497, AY954886,
786 HM116385, DQ650344, KY490554, KC305650, EF547155, FJ769050, HM243547, and
787 AY524773 accession codes.
788 The *gp60* sequence data that supported the CryptoCap_100k bait design were
789 from GenBank with the FJ490060, FJ490087, FJ490058, JX412915, JX412926,
790 JX412925, KC204983, and KC204984 accession codes.
791 The twelve *Cryptosporidium* genome sequences used for CryptoCap_100k bait
792 set characterization and *in silico* simulations were from GenBank with the
793 GCF_00000165345.1, GCA_004337835.1, GCA_004337795.1, GCA_007210665.1,
794 GCF_001865345.1, GCA_001865355.1, GCA_001593455.1, GCF_000006515.1,
795 GCA_000165345.1, GCA_001483515.1, and GCA_002223825.1 accession codes *C.*
796 *hominis* UdeA01 and *C. parvum* IOWA-ATCC were removed for those analyses with
797 only ten genome sequences reference database (10-CrypGS).
798 Generated fastq files from all *in silico* and *in vitro* experiments for mock and
799 clinical samples are available in GenBank BioProject number PRJNA1061798. Clinical

samples were scrubbed of human DNA upon submission by NCBI using their Human Read Removal Tool.

Computer Code

The code and data used for all sections of this paper are described in the Supplementary Notes, also available in Dryad, doi: 10.5061/dryad.gtht76j0h.

References

1. Hunter, P. R. & Nichols, G. Epidemiology and clinical features of *Cryptosporidium* infection in immunocompromised patients. *Clin Microbiol Rev* **15**, 145–154 (2002).
2. Khan, A., Shaik, J. S. & Grigg, M. E. Genomics and molecular epidemiology of *Cryptosporidium* species. *Acta Trop* **184**, 1–14 (2018).
3. Wang, J. *et al.* Long-term treatment with green tea polyphenols modifies the gut microbiome of female sprague-dawley rats. *J Nutr Biochem* **56**, 55–64 (2018).
4. Bouzid, M., Hunter, P. R., Chalmers, R. M. & Tyler, K. M. *Cryptosporidium* pathogenicity and virulence. *Clin Microbiol Rev* **26**, 115–134 (2013).
5. Chalmers, R. M. & Katzer, F. Looking for *Cryptosporidium*: the application of advances in detection and diagnosis. *Trends Parasitol* **29**, 237–251 (2013).
6. 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. *Lancet Glob Health* **6**, e758–e768 (2018).
7. Gilbert, I. H. *et al.* Safe and effective treatments are needed for cryptosporidiosis, a truly neglected tropical disease. *BMJ Global Health* **8**, e012540 (2023).
8. Sow, S. O. *et al.* The Burden of *Cryptosporidium* diarrheal disease among children < 24 months of age in moderate/high mortality regions of Sub-Saharan Africa and South Asia, utilizing data from the Global Enteric Multicenter Study (GEMS). *PLOS Neglected Tropical Diseases* **10**, e0004729 (2016).
9. GBD Results. *Institute for Health Metrics and Evaluation* https://vizhub.healthdata.org/gbd-results.

10. 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. *The Lancet* **382**, 209–222 (2013).
11. Bhutta, Z. A. & Black, R. E. Global maternal, newborn, and child health--so near and yet so far. *N Engl J Med* **369**, 2226–2235 (2013).
12. Cryptosporidiosis NNDSS Summary Report for 2019 | Water-related Topics | Healthy Water | CDC. https://www.cdc.gov/healthywater/surveillance/cryptosporidium/cryptosporidium-2019.html (2021).
13. Feng, Y., Ryan, U. M. & Xiao, L. Genetic diversity and population structure of *Cryptosporidium*. *Trends Parasitol* **34**, 997–1011 (2018).
14. Morris, A., Robinson, G., Swain, M. T. & Chalmers, R. M. Direct sequencing of *Cryptosporidium* in stool samples for public health. *Front Public Health* **7**, 360 (2019).
15. Ryan, U. M., Feng, Y., Fayer, R. & Xiao, L. Taxonomy and molecular epidemiology of *Cryptosporidium* and *Giardia* – a 50 year perspective (1971–2021). *International Journal for Parasitology* **51**, 1099–1119 (2021).
16. Nader, J. L. *et al.* Evolutionary genomics of anthroponosis in *Cryptosporidium*. *Nat Microbiol* **4**, 826–836 (2019).
17. Baptista, R. P., Cooper, G. W. & Kissinger, J. C. Challenges for *Cryptosporidium* Population Studies. *Genes (Basel)* **12**, 894 (2021).
18. Baptista, R. P. *et al.* Long-read assembly and comparative evidence-based reanalysis of *Cryptosporidium* genome sequences reveal expanded transporter repertoire and

duplication of entire chromosome ends including subtelomeric regions. *Genome Res* **32**, 203–213 (2022).

19. Šlapeta, J. Cryptosporidiosis and Cryptosporidium species in animals and humans: a thirty colour rainbow? *Int J Parasitol* **43**, 957–970 (2013).
20. Ruecker, N. J., Hoffman, R. M., Chalmers, R. M. & Neumann, N. F. Detection and resolution of *Cryptosporidium* species and species mixtures by genus-specific nested PCR-restriction fragment length polymorphism analysis, direct sequencing, and cloning. *Appl Environ Microbiol* **77**, 3998–4007 (2011).
21. Wang, Y. *et al.* Population genetics of *Cryptosporidium meleagridis* in humans and birds: evidence for cross-species transmission. *International Journal for Parasitology* **44**, 515–521 (2014).
22. Yang, Z. *et al.* Multilocus sequence typing and population genetic structure of *Cryptosporidium cuniculus* in rabbits in Heilongjiang Province, China. *Infection, Genetics and Evolution* **64**, 249–253 (2018).
23. Tang, Y. *et al.* Development of a multilocus sequence typing tool for high-resolution subtyping and genetic structure characterization of *Cryptosporidium ubiquitum*. *Infection, Genetics and Evolution* **45**, 256–261 (2016).
24. Uran-Velasquez, J. *et al.* Multilocus Sequence Typing helps understand the genetic diversity of *Cryptosporidium hominis* and *Cryptosporidium parvum* isolated from Colombian patients. *PLOS ONE* **17**, e0270995 (2022).
25. Hawash, Y. DNA extraction from protozoan oocysts/cysts in feces for diagnostic PCR. *Korean J Parasitol* **52**, 263–271 (2014).

26. Corsi, G. I. *et al.* Recent genetic exchanges and admixture shape the genome and population structure of the zoonotic pathogen Cryptosporidium parvum. *Molecular Ecology* **32**, 2633–2645 (2023).

27. Tichkule, S. *et al.* Global population genomics of two subspecies of *Cryptosporidium* hominis during 500 years of evolution. *Molecular Biology and Evolution* **39**, msac056 (2022).

28. Gan, M., Liu, Q., Yang, C., Gao, Q. & Luo, T. Deep whole-genome sequencing to detect mixed infection of *Mycobacterium tuberculosis*. *PLOS ONE* **11**, e0159029 (2016).

29. Kaupke, A., Gawor, J., Rzeżutka, A. & Gromadka, R. Identification of pig-specific *Cryptosporidium* species in mixed infections using Illumina sequencing technology. *Exp Parasitol* **182**, 22–25 (2017).

30. Widmer, G. *et al.* Population structure of natural and propagated isolates of *Cryptosporidium parvum*, *C. hominis* and *C. meleagridis*. *Environmental Microbiology* **17**, 984–993 (2015).

31. Guo, Y. *et al.* Isolation and enrichment of *Cryptosporidium* DNA and verification of DNA purity for whole-genome sequencing. *Journal of Clinical Microbiology* **53**, 641–647 (2015).

32. Guo, Y. *et al.* Comparative genomic analysis reveals occurrence of genetic recombination in virulent *Cryptosporidium* hominis subtypes and telomeric gene duplications in *Cryptosporidium parvum*. *BMC Genomics* **16**, 320 (2015).

33. Hadfield, S. J. *et al.* Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples. *BMC Genomics* **16**, 650 (2015).

34. Ruecker, N. J. *et al.* Tracking host sources of *Cryptosporidium* spp. in raw water for improved health risk assessment. *Appl Environ Microbiol* **73**, 3945–3957 (2007).

35. Cama, V. *et al.* Mixed *Cryptosporidium* infections and HIV. *Emerg Infect Dis* **12**, 1025–1028 (2006).

36. Jiang, J., Alderisio, K. A. & Xiao, L. Distribution of *Cryptosporidium* genotypes in storm event water samples from three watersheds in New York. *Appl Environ Microbiol* **71**, 4446–4454 (2005).

37. McLauchlin, J., Amar, C., Pedraza-Díaz, S. & Nichols, G. L. Molecular epidemiological analysis of *Cryptosporidium* spp. in the United Kingdom: results of genotyping *Cryptosporidium* spp. in 1,705 fecal samples from humans and 105 fecal samples from livestock animals. *J Clin Microbiol* **38**, 3984–3990 (2000).

38. Ng-Hublin, J. S. Y., Combs, B., Mackenzie, B. & Ryan, U. Human cryptosporidiosis diagnosed in Western Australia: a mixed infection with *Cryptosporidium meleagridis*, the *Cryptosporidium* mink genotype, and an unknown *Cryptosporidium* species. *J Clin Microbiol* **51**, 2463–2465 (2013).

39. Sharma, P., Sharma, A., Sehgal, R., Malla, N. & Khurana, S. Genetic diversity of *Cryptosporidium* isolates from patients in North India. *Int J Infect Dis* **17**, e601-605 (2013).

40. Reed, C., Sturbaum, G. D., Hoover, P. J. & Sterling, C. R. *Cryptosporidium parvum* mixed genotypes detected by PCR-restriction fragment length polymorphism analysis. *Appl Environ Microbiol* **68**, 427–429 (2002).

41. Korpe, P. S. *et al.* Case-control study of *Cryptosporidium* transmission in Bangladeshi households. *Clin Infect Dis* **68**, 1073–1079 (2019).

42. Tanriverdi, S., Arslan, M. Ö., Akiyoshi, D. E., Tzipori, S. & Widmer, G. Identification of genotypically mixed *Cryptosporidium parvum* populations in humans and calves. *Molecular and Biochemical Parasitology* **130**, 13–22 (2003).

43. Mallon, M. *et al.* Population structures and the role of genetic exchange in the zoonotic pathogen *Cryptosporidium parvum*. *J Mol Evol* **56**, 407–417 (2003).

44. Gaudin, M. & Desnues, C. Hybrid capture-based next generation sequencing and its application to human infectious diseases. *Frontiers in Microbiology* **9**, (2018).

45. Glenn, T. C. & Faircloth, B. C. Capturing Darwin's dream. *Mol Ecol Resour* **16**, 1051–1058 (2016).

46. Quek, Z. B. R. & Ng, S. H. Hybrid-capture target enrichment in human pathogens: identification, evolution, biosurveillance, and genomic epidemiology. *Pathogens* **13**, 275 (2024).

47. Khan, A. *et al.* Phylogenomic reconstruction of *Cryptosporidium* spp. captured directly from clinical samples reveals extensive genetic diversity. 2024.04.17.589752 Preprint at https://doi.org/10.1101/2024.04.17.589752 (2024).

48. Abrahamsen, M. S. *et al.* Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* **304**, 441–445 (2004).

49. Robinson, G., Elwin, K. & Chalmers, R. M. *Cryptosporidium* diagnostic assays: molecular detection. in *Cryptosporidium: Methods and Protocols* (eds Mead, J. R. & Arrowood, M. J.) 11–22 (Springer, New York, NY, 2020). doi:10.1007/978-1-4939-9748-0_2.

50. Baptista, R. de P., Xiao, R., Li, Y., Glenn, T. C. & Kissinger, J. C. New T2T assembly of *Cryptosporidium parvum* IOWA II annotated with legacy-compatible gene identifiers. *Sci Data* **12**, 1039 (2025).

51. Simon, M. D. & Machyna, M. Principles and practices of hybridization capture experiments to study long noncoding RNAs that act on chromatin. *Cold Spring Harb Perspect Biol* **11**, a032276 (2019).

52. Dąbrowska, J., Sroka, J. & Cencek, T. Investigating *Cryptosporidium* spp. using genomic, proteomic and transcriptomic techniques: current progress and future directions. *International Journal of Molecular Sciences* **24**, 12867 (2023).

53. Robinson, G. *et al.* Re-description of *Cryptosporidium cuniculus* Inman and Takeuchi, 1979 (Apicomplexa: Cryptosporidiidae): Morphology, biology and phylogeny. *International Journal for Parasitology* **40**, 1539–1548 (2010).

54. Chalmers, R. M., Elwin, K., Hadfield, S. J. & Robinson, G. Sporadic Human Cryptosporidiosis Caused by *Cryptosporidium cuniculus*, United Kingdom, 2007–2008 -. *Emerging Infectious Diseases journal - CDC* **17**, (2011).

55. Xiao, L. Molecular epidemiology of cryptosporidiosis: an update. *Exp Parasitol* **124**, 80–89 (2010).

56. Agyabeng-Dadzie, F., Xiao, R. & Kissinger, J. C. *Cryptosporidium* Genomics — current understanding, advances, and applications. *Curr Trop Med Rep* **11**, 92–103 (2024).

57. Gatei, W. *et al.* Multilocus sequence typing and genetic structure of *Cryptosporidium hominis* from children in Kolkata, India. *Infect Genet Evol* **7**, 197–205 (2007).

58. Hadfield, S. J., Robinson, G., Elwin, K. & Chalmers, R. M. Detection and differentiation of *Cryptosporidium* spp. in human clinical samples by use of real-time PCR. *J Clin Microbiol* **49**, 918–924 (2011).

59. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

61. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

62. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165 (2016).

63. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).

64. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).

65. Paparini, A. *et al.* Comparison of Sanger and next generation sequencing performance for genotyping *Cryptosporidium* isolates at the 18S rRNA and actin loci. *Exp Parasitol* **151–152**, 21–27 (2015).

66. Puiu, D., Enomoto, S., Buck, G. A., Abrahamsen, M. S. & Kissinger, J. C. CryptoDB: the *Cryptosporidium* genome resource. *Nucleic Acids Res* **32**, D329-331 (2004).

67. Bushnell, B. BBMap: a fast, accurate, splice-aware aligner. https://escholarship.org/uc/item/1h3515gn (2014).

68. Lindgreen, S., Adair, K. L. & Gardner, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci Rep* **6**, 19233 (2016).

69. Glenn, T. C. *et al*. Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext). *PeerJ* **7**, e7755 (2019).

70. Stroup, S. E. *et al*. Real-time PCR detection and speciation of *Cryptosporidium* infection using Scorpion probes. *J Med Microbiol* **55**, 1217–1222 (2006).

71. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res* **22**, 939–946 (2012).

72. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

73. Penumarthi, L. R., Baptista, R. P., Beaudry, M. S., Glenn, T. C. & Kissinger, J. C. A new chromosome-level genome assembly and annotation of Cryptosporidium meleagridis. *bioRxiv* 2024.02.16.580748 (2024) doi:10.1101/2024.02.16.580748.

74. Kuhn, M. & Davis, V. parsnip: a common API to modeling and analysis functions. (2025).

75. McKenna, A. *et al*. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**, 1297–1303 (2010).

76. Knaus, B. J. & Grünwald, N. J. vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources* **17**, 44–53 (2017).

77. Mijangos, J. L., Gruber, B., Berry, O., Pacioni, C. & Georges, A. dartR v2: an accessible genetic analysis platform for conservation, ecology and agriculture. *Methods in Ecology and Evolution* **13**, 2150–2158 (2022).

78. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

79. Dray, S. & Dufour, A.-B. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software* **22**, 1–20 (2007).

80. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

81. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).

82. Kamvar, Z. N., Tabima, J. F. & Grünwald, N. J. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281 (2014).

83. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

84. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

85. Pina-Martins, F., Silva, D. N., Fino, J. & Paulo, O. S. Structure_threader: An improved method for automation and parallelization of programs structure, fastStructure and

MavericK on multicore CPU systems. *Molecular Ecology Resources* **17**, e268–e274 (2017).

86. Earl, D. A. & vonHoldt, B. M. Structure Harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genet Resour* **4**, 359–361 (2012).
87. Wickham, H., Hester, J. & Bryan, J. readr: read rectangular text data. R package version 2.1.5. *https://github.com/tidyverse/readr* https://github.com/tidyverse/readr (2024).
88. Tonkin-Hill, G. & Lee, S. Starmie: population structure model inference and visualisation. (2016).
89. Goudet, J. hierfstat, a package for r to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**, 184–186 (2005).
90. Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software* **3**, 731 (2018).
91. Wilke, C. ggridges: ridgeline Plots in 'ggplot2'. (2024).
92. Zhu, S. J., Almagro-Garcia, J. & McVean, G. Deconvolution of multiple infections in *Plasmodium falciparum* from high throughput sequencing data. *Bioinformatics* **34**, 9–15 (2018).
93. Danecek, P. *et al*. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
94. Lenth, R. V. Least-squares means: the R package lsmeans. *Journal of Statistical Software* **69**, 1–33 (2016).

95. Wickham, H. *et al*. Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

96. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

97. Plotly Technologies Inc. Collaborative data science. https://plot.ly (2015).

Tables

Table 1. Summary of *Cryptosporidium* species considered, genome version used, and criteria for designing CryptoCap_100k bait set

Identifier	Reference Genome	Criteria (Identity: Overlap)	Number of Baits
Whole Genome			
C. parvum	GCA_015245375.1	Reference	73,672
C. hominis	GCA_001483515.1	90%:50%	2,851
C. cuniculus	GCA_004337835.1	90%:50%	738
C. tyzzeri	GCA_007210665.1	90%:50%	858
C. meleagridis	GCA_001593445.1	85%:45%	9,034
C. viatorum	GCA_004337795.1	85%:45%	11,774
***C. viatorum* (select genes)**		50%:95%	943
18S rRNA		50%:95%	63
gp60		NA	67
		Total Number of Baits	100,000

GC Content	30.1%
Soft Masked	2.8% (baits ranged from 0–50% masked)

Table 2. Summary statistics from enriching 100 human fecal DNA samples with two library preparation methods using CryptoCap_100k baits

	iTru Libraries			iNextEra Libraries		
Average values	Not enriched	Single enriched		Not enriched	Single enriched	Double enriched
% retained after quality filtering	61.5	83.9		89.7	90.8	94.5
% reads on target *C. parvum*	0.04	35.6		0.03	15.8	65.0
% reads on target 10-CrypGS	0.1	41.2		0.10	17.9	65. 8
Enrichment factor [1]	N/A	2,095		N/A	1,112	10,700[2]
Depth per 100k [1]	0.00	1.12		0.001	0.563	2.702
Breadth per 100k[1]	0.00	2.25		0.065	1.79	5.99
Number of samples	73	100		91	91	91

[1] matching *C. parvum*)

[2] This value corresponds to the enrichment factor of double-enriched samples compared to non-enriched counterparts, which is not displayed in Tables S14 and S15.

Figures

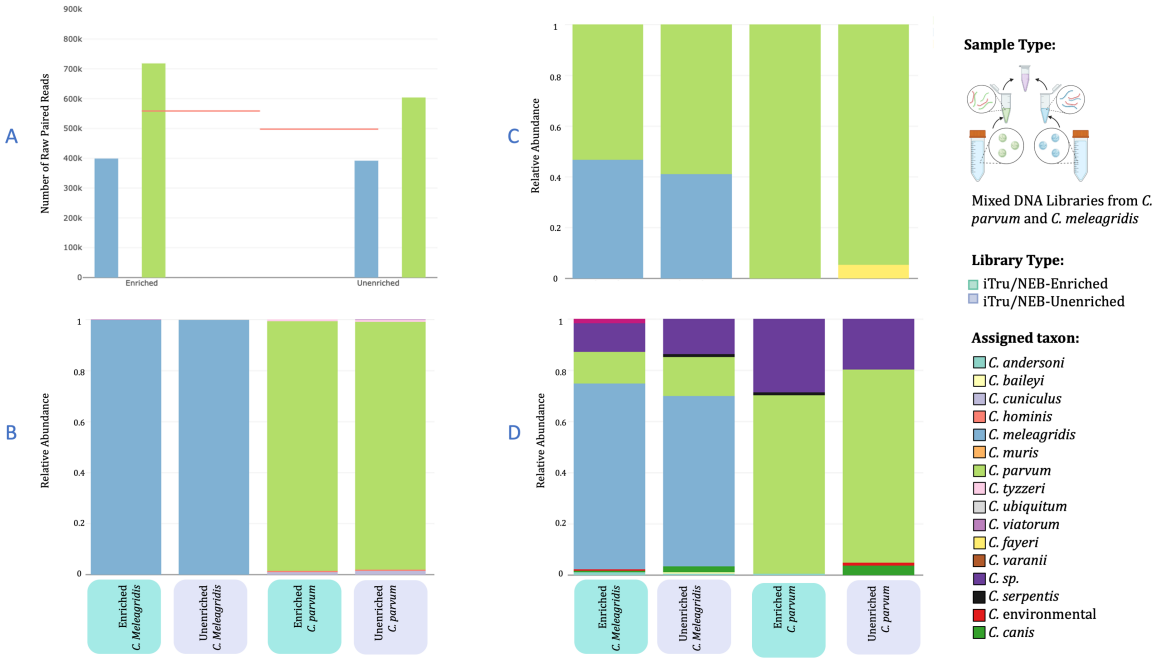


Figure 1. Species assignment plots based on mapping reads from unenriched and enriched libraries for one mock mixed-infection sample. (A) The number of raw paired reads obtained for each species in the mix, *C. parvum* in green and *C. meleagridis* in blue, for both library types. The pink lines indicate the expected number of reads for each sample for a 50-50 proportion. Relative abundance of species assigned from mapping the sequence reads to (B) the ten-genome sequence reference database (10-CrypGS), (C) the *gp60* database, and (D) the *18S rRNA* database.

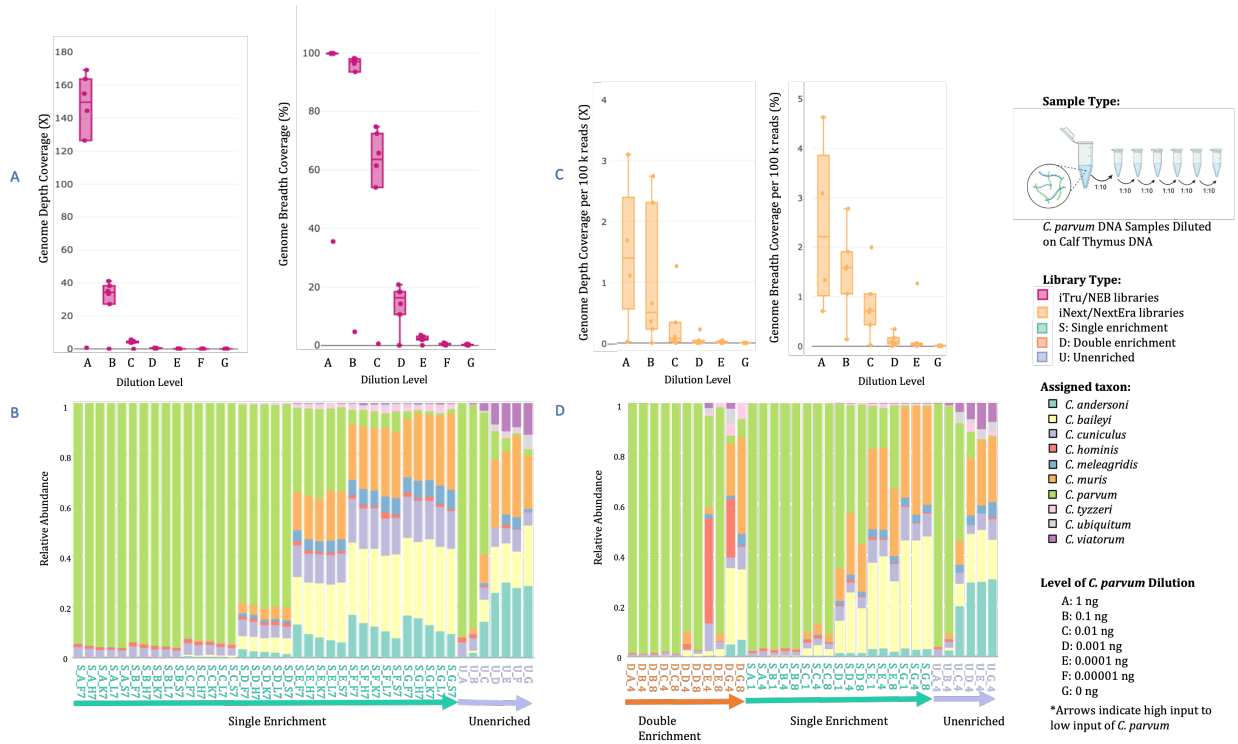


Figure 2. Summary statistics for sequence data from serially diluted *C. parvum* samples to assess detection levels given different input amounts. (A) *C. parvum* genome depth and breadth of coverage for diluted DNA samples prepared with NEB/ iTru libraries (B) species assignment of diluted *C. parvum* samples prepared with NEB-iTru protocol and unenriched (*U*) or single-enriched (*S*) mapped to the ten *Cryptosporidium* genome sequences reference (10-CrypGS); (C) *C. parvum* genome depth and breadth coverage for diluted DNA samples prepared with iNext/NextEra libraries, (D) species assignment of diluted *C. parvum* samples prepared with iNextEra protocol and unenriched (*U*), single-enriched (*S*), or double-enriched (*D*) mapped to the ten *Cryptosporidium* genome sequences reference (10-CrypGS). In the species assignment plots, the first letter in the sample name corresponds to the library type, and the second letter corresponds to the level of input of DNA (A–G). The direction of the arrow indicates high to low input (A → G).

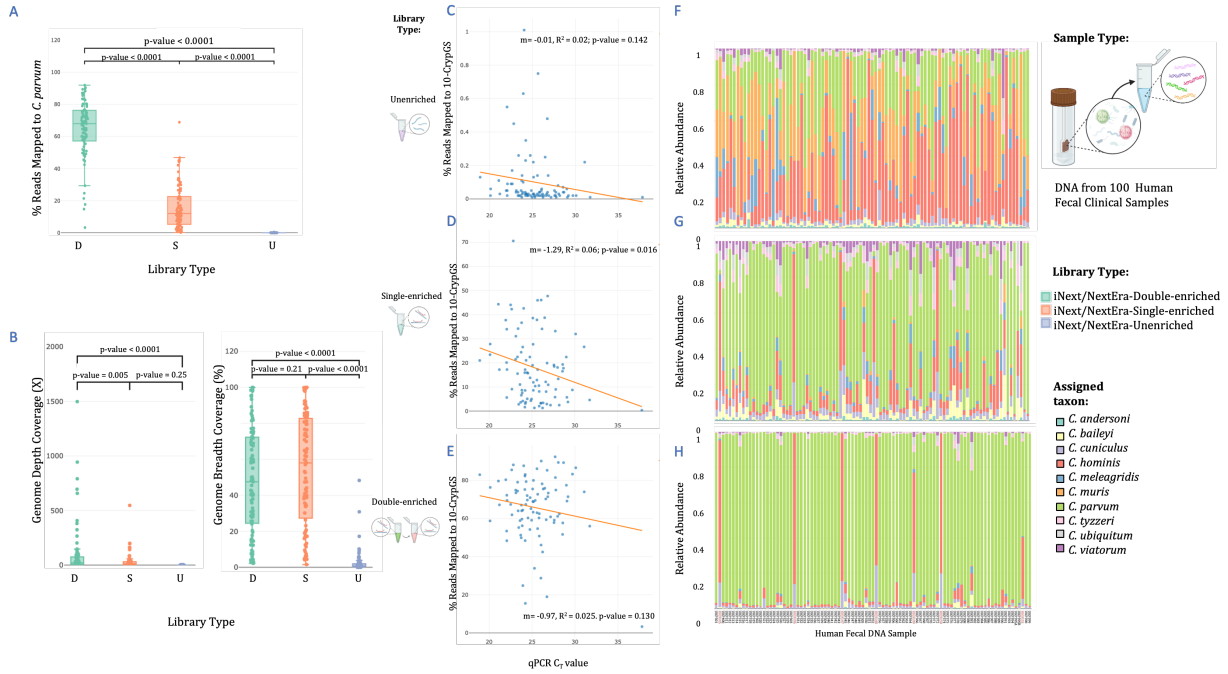


Figure 3. Summary statistics for sequence data from 91 clinical samples prepared with the iNextEra protocol and iNext primers, comparing libraries enriched with CryptoCap_100k bait set versus unenriched libraries for validation. (A) Boxplot of percentage of reads mapped to *C. parvum* from double-enriched [D], single-enriched [S], and unenriched libraries[U]. (B) Genome depth and breadth of coverage for each sample in double-enriched, single-enriched, and unenriched libraries. Linear regression model for predicting the percentage of mapped reads to a reference database containing ten *Cryptosporidium* genome sequences (10-CrypGS) given a C_T value from qPCR for (C) unenriched libraries and (D) single-enriched libraries with CryptoCap_100k, and (E) double-enriched libraries with CryptoCap_100k. Species assignment of each sample prepared with iNextEra protocol with (F) no further enrichment, (G) single-enriched with CryptoCap_100k, and (H) double-enriched with CryptoCap_100k; when mapped to a reference database containing ten

Cryptosporidium genome sequences (10-CrypGS). Sample names for plot 4F are not shown.

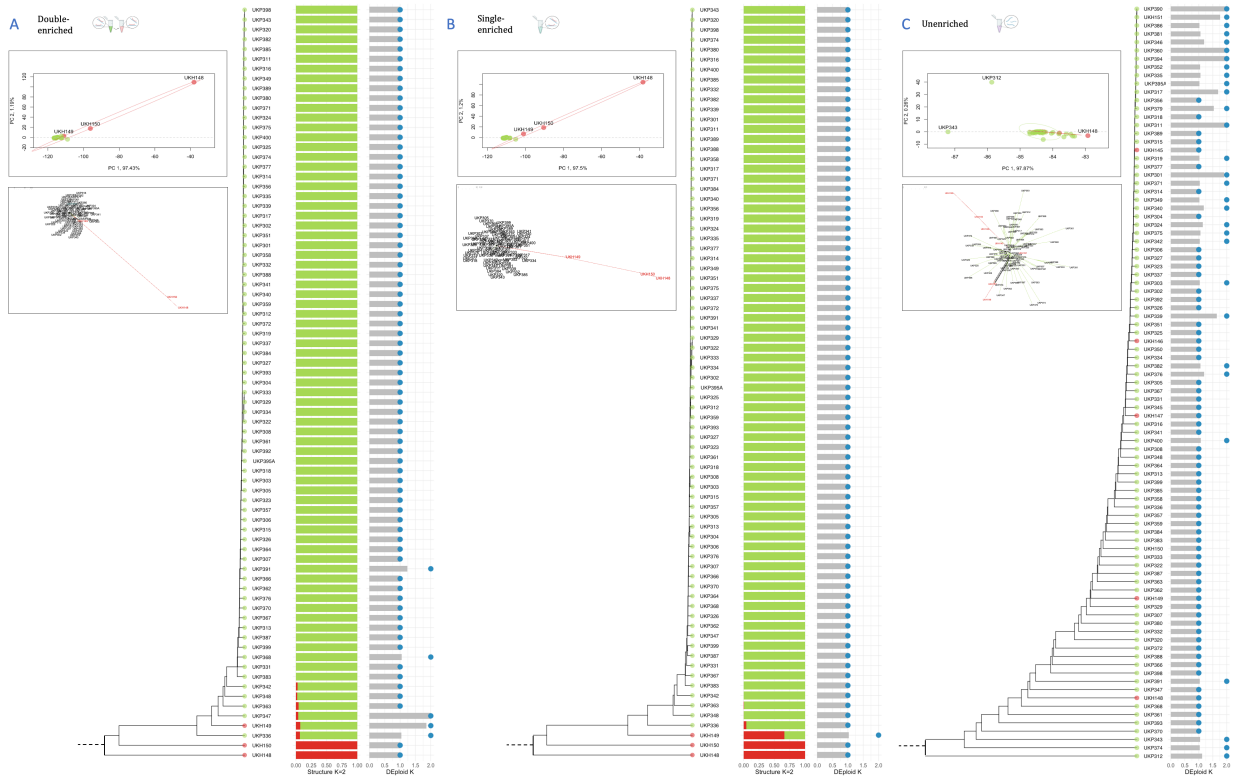


Figure 4. Population genomic analyses from called SNP variants of clinical samples prepared with the iNextEra protocol and iNext primers. Libraries were double-, single-enriched, and unenriched, respectively (A, B, and C). Each panel includes: Principal Components Analyses, phylogenetic distance networks, UPGMA Phylogenetic trees of Nei's genetic distances, assignment plots from Structure analyses for K=2 (selected based on Evanno's method—unenriched samples could not be analyzed due to high levels of missing data), and effective (gray bars) and inferred (blue circles) number of strains (K) detected in each sample by the DEploid software.

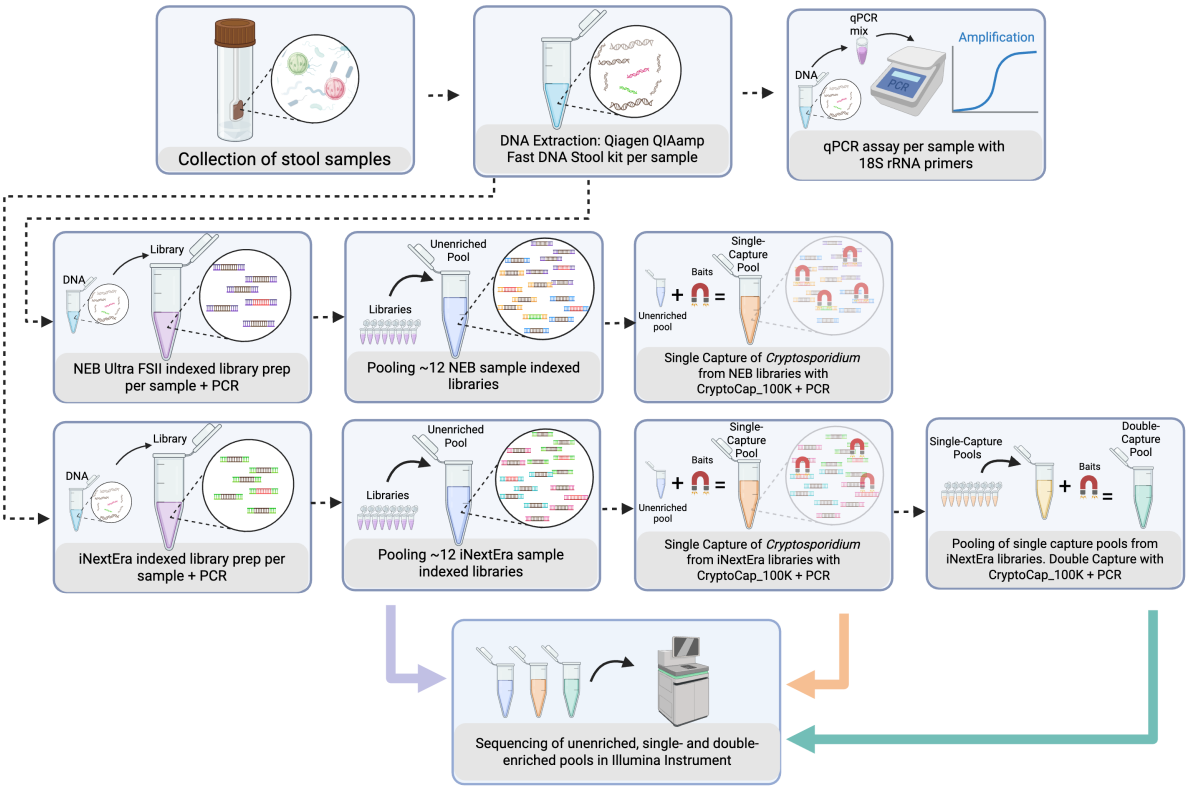
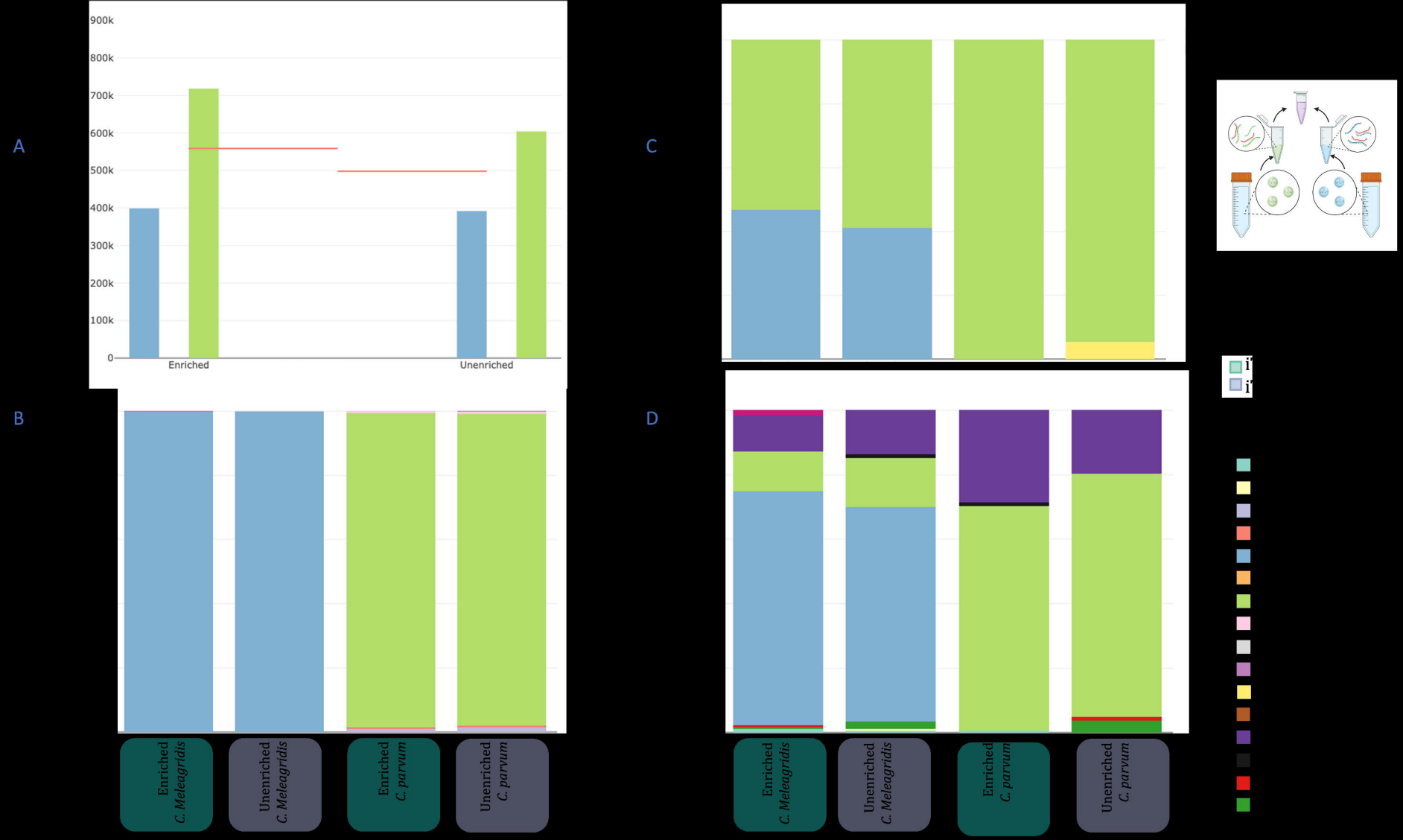
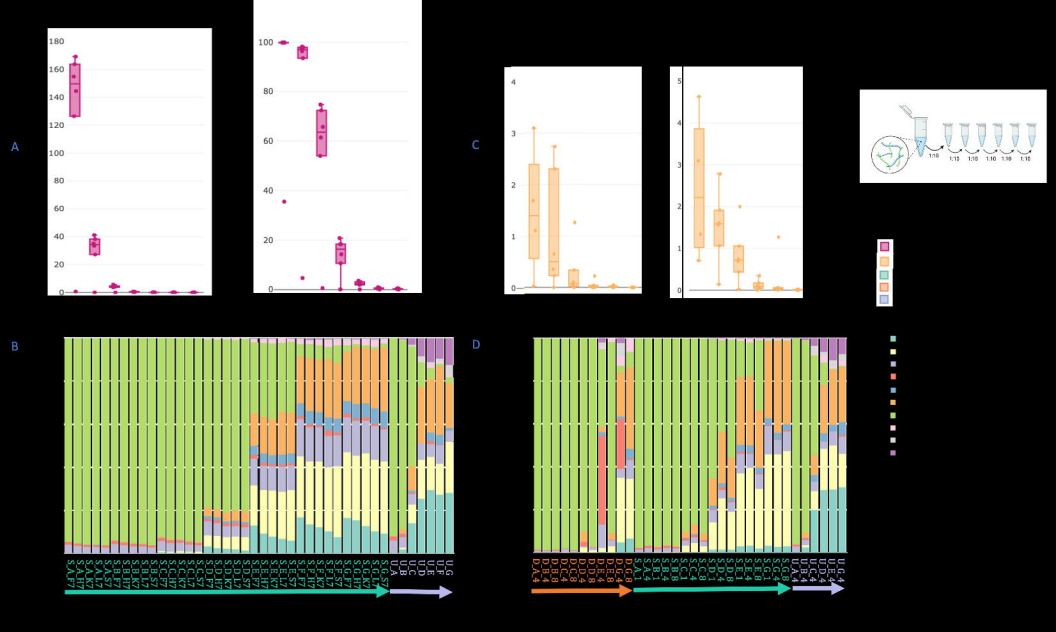


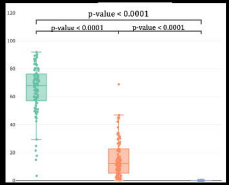
Figure 5. Laboratory workflow used for the analysis of a set of clinical samples from the UK. The genomic DNA of each patient's sample was extracted, and *Cryptosporidium* was validated via qPCR. Each DNA extract was used for two types of library prep, NEB Ultra FSII and iNextEra library prep. Unenriched library preparation aliquots were used for sequencing. Also, aliquots from unenriched libraries were pooled according to library type, and enrichments were performed with CryptoCap_100k on these pools. Aliquots of these products, referred to as single-enriched products, were used for sequencing. Aliquots of single-enriched libraries, which used the iNextEra protocol, were pooled, and enrichments were performed again with CryptoCap_100k. Aliquots of these products, named double-enriched products, were used for sequencing. Colors represent those used throughout the manuscript to differentiate unenriched,

single-enriched, and double-enriched products. Created in BioRender. Bayona N (2025) https://BioRender.com/yysbjqh.

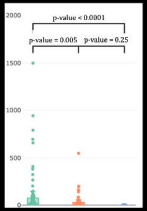

A
900k
800k
700k
600k
500k
400k
300k
200k
100k
0
Enriched
Unenriched
B
Enriched C. Meleagridis
Unenriched C. Meleagridis
Enriched C. parvum
Unenriched C. parvum
C
D
Enriched C. Meleagridis
Unenriched C. Meleagridis
Enriched C. parvum
Unenriched C. parvum

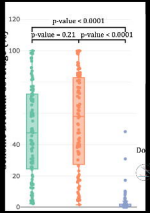

A
B
C
D
1:10
1:10
1:10
1:10
1:10

A

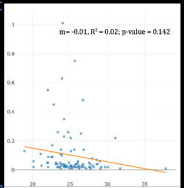

p-value < 0.0001
p-value < 0.0001
p-value < 0.0001

B

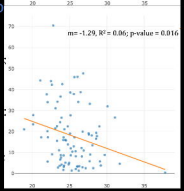

p-value < 0.0001
p-value = 0.005
p-value = 0.25


p-value < 0.0001
p-value = 0.21
p-value < 0.0001

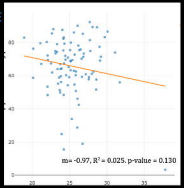
C


m= -0.01, R² = 0.02; p-value = 0.142

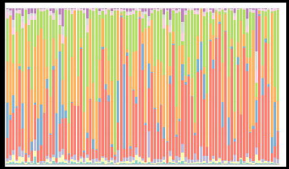
D

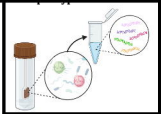

m= -1.29, R² = 0.06; p-value = 0.016

E

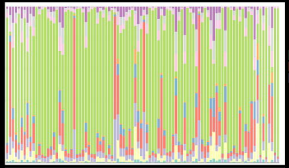

m= -0.97, R² = 0.025, p-value = 0.130

F

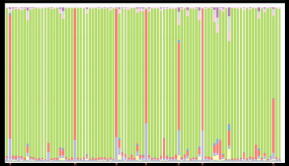


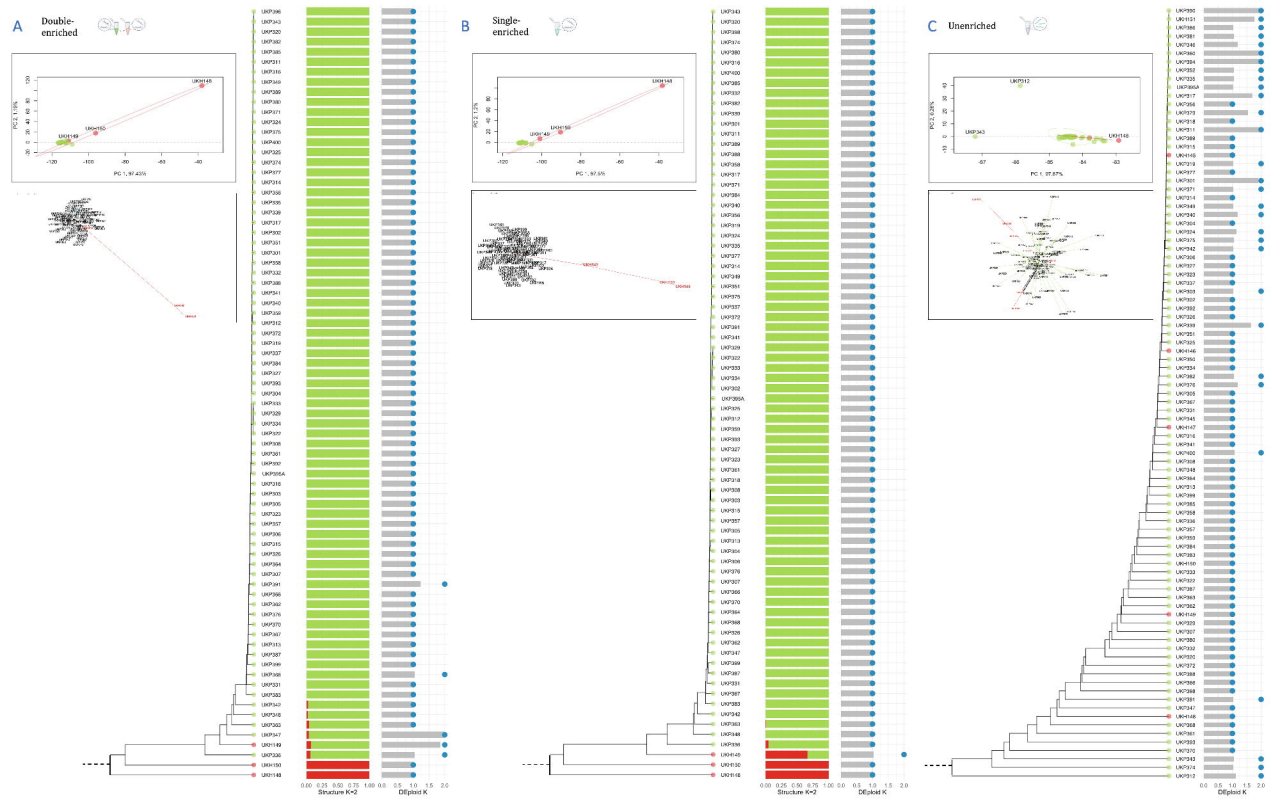


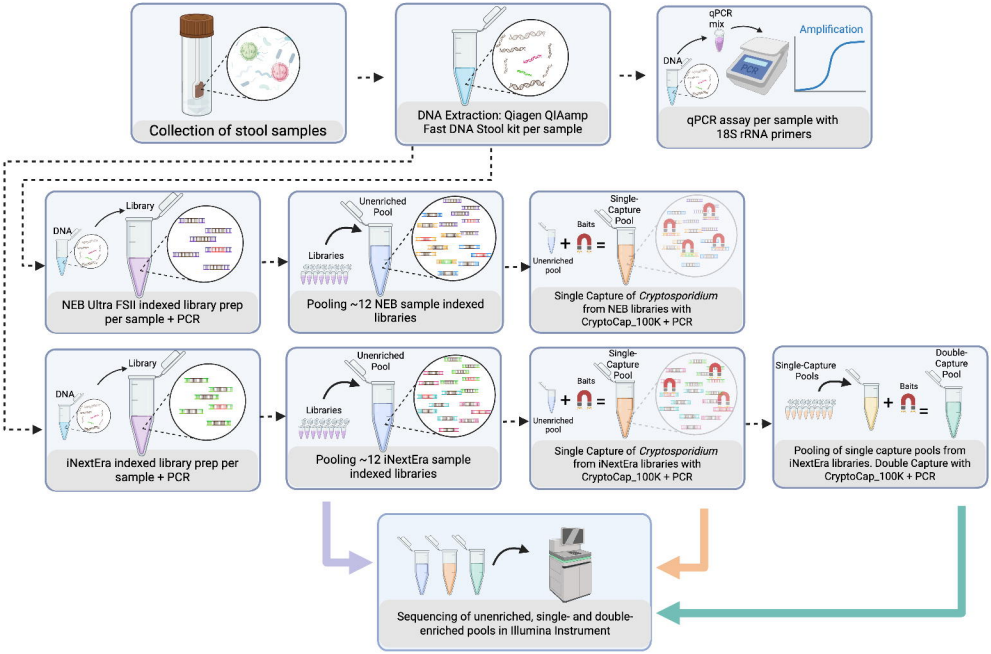
G



H




A
Double-enriched
B
Single-enriched
C
Unenriched
PC 1, 97.43%
PC 2, 1.15%
PC 1, 97.5%
PC 2, 0.25%
PC 1, 97.87%
Structure K=2
DEploid K


Collection of stool samples
DNA Extraction: Qiagen QIAamp Fast DNA Stool kit per sample
qPCR mix
DNA
Amplification
qPCR assay per sample with 18S rRNA primers
Library
DNA
NEB Ultra FSII indexed library prep per sample + PCR
Unenriched Pool
Libraries
Pooling ~12 NEB sample indexed libraries
Single-Capture Pool
Baits
Unenriched pool
Single Capture of Cryptosporidium from NEB libraries with CryptoCap_100K + PCR
Library
DNA
iNextEra indexed library prep per sample + PCR
Unenriched Pool
Libraries
Pooling ~12 iNextEra sample indexed libraries
Single-Capture Pool
Baits
Unenriched pool
Single Capture of Cryptosporidium from iNextEra libraries with CryptoCap_100K + PCR
Single-Capture Pools
Double-Capture Pool
Baits
Pooling of single capture pools from iNextEra libraries. Double Capture with CryptoCap_100K + PCR
Sequencing of unenriched, single- and double-enriched pools in Illumina Instrument