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 human-infecting *Cryptosporidium* species (*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* spp. 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 the reliability of detecting species and mixed infections within a sample, and allows assessment of genetic variation via SNP calling, while decreasing costs.

47 **Introduction**

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

58 There are at least 39 recognized species within the genus *Cryptosporidium*[13,14], with a
59 wide range of host specificity and likelihood of infecting humans. Thus, *Cryptosporidium*
60 spp. vary in public health importance[13]. At the genome sequence level, a cluster of
61 species shares greater than 80% identity across much of their genomes[15–18], including *C.*
62 *cuniculus*, *C. hominis*, *C. meleagridis*, *C. parvum*, *C. tyzzeri,* and *C. viatorum*.

63 Several molecular methods have been developed to gain insight into the genetic
64 diversity of *Cryptosporidium*, often relying on single-marker typing of conserved (e.g.,
65 rRNA-SSU) or polymorphic genes (e.g., *gp60* or *hsp70*)[2,19]. Other approaches include
66 multi-locus sequence typing (MLST)[20–23] and, more recently, whole-genome sequencing
67 (WGS). The disadvantages of MLST include unreliability of markers in different regions
68 due to limited information about the genetic variability of *Cryptosporidium*[23], and thus,
69 assays are not universal. Challenges with methods such as WGS are constrained by the

70 need to purify oocysts[24], given the limited amount of target DNA from fecal clinical
71 samples (which also have an abundance of other organisms' DNA) [25,26]. These challenges
72 have led to the use of purification methods combined with costly deep sequencing[27,28].
73 For example, to obtain sufficient DNA, oocysts may be propagated through animals,
74 which has been shown to change *Cryptosporidium* populations through selection of
75 individual subtypes or recombination[14,29]. Some WGS analyses have relied on whole-
76 genome amplification protocols to overcome low concentrations of target DNA[14,25,26,30].
77 A helpful alternative that still requires semi-purified oocyst suspensions[4,14,31], and that
78 may be prone to uneven amplification and the production of chimeric DNA fragments.
79 Other promising protocols for WGS directly from human stool samples do not require
80 whole-genome amplification but rely on a sufficient sample volume and an adequate
81 concentration of oocysts, achieving WGS data for ~53% of samples[32], showing medium
82 success rates. Thus, due to the limitations of existing approaches, additional methods that
83 can characterize *Cryptosporidium* genomes from limited amounts of DNA, if possible,
84 directly from fecal samples, which can be implemented easily and reliably, would be
85 beneficial.
86 Host-derived and environmental samples can also contain more than one species or
87 genotype of *Cryptosporidium* spp.[33–38]. Mixed infections may be limited to specific
88 geographic areas, particularly in endemic regions[39]. Most reported human mixed-species
89 infections contain *C. parvum* and *C. hominis*[34,36,40]; still, mixed infections are more
90 commonly represented by multiple strains of a single species[16]. Their prevalence has
91 been reported to range from 5% in the US[41] to 12% in Scotland[42], and up to 33% in
92 Peru[34]. However, their frequency and prevalence by different species or strains are

93 challenging to estimate, given the reliance on limited reference databases and PCR-
94 generated data that may fail to detect low-abundance types/species within mixed
95 infection samples[19]. Thus, methods that enable the detection and characterization of
96 mixed infections are necessary to gain a better understanding of cryptosporidiosis.

97 Hybridization sequence capture has well-known advantages[43–45] relevant to studying
98 *Cryptosporidium* spp. and other parasites: it is feasible to design thousands of baits to
99 capture genomes, such as the 9 Mb genome of *Cryptosporidium* spp.; baits can tolerate
100 mismatches (thus regions with < 20% sequence divergence may be captured); and
101 successful enrichment reduces the amount of sequencing needed (thereby reducing
102 sequencing costs in comparison with non-enriched samples). The first *Cryptosporidium*
103 whole-genome bait set (CryptoCap_75k[46]) was designed using the *C. parvum* IOWAII
104 incomplete genome assembly[47], and is reported to perform well when each sample is
105 enriched individually. However, library pools that contain diverse species showed bias
106 during enrichment for *C. parvum* over other distantly related species[46]. Because fecal
107 samples can contain unknown or mixed *Cryptosporidium* species, a second-generation
108 bait set is needed to enrich genomes for a broader spectrum of *Cryptosporidium* species
109 and consider highly polymorphic genomic regions. Enriching pools of sample libraries,
110 rather than single libraries, is also desirable for time- and cost-efficiency.

111 To ameliorate the current challenges in the genomic study of human-infecting
112 *Cryptosporidium* spp., we developed, validated, and critically assessed CryptoCap_100k.
113 Motivated by CryptoCap_75k[46], we advance this approach by creating a bait set that also
114 captures highly diverse genomic regions (e.g., including all subtelomeric regions) and a
115 wide diversity of human-infecting species. CryptoCap_100k was designed (without

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

126 **Results**

127 *CryptoCap_100k Bait Design & Characteristics*

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

133 Mapping the baits to twelve *Cryptosporidium* genome sequences (Table 2,
134 Supplementary Notes #1) revealed ~90.46% breadth of genome coverage at 1X for target
135 species, but much lower values for distant species not considered in bait design (7.33% at
136 0.08X; Table 2; Fig. S1).

137 When comparing the multi-species design (CryptoCap_100k) to the single-species
138 design (CryptoCap_75k) (Supplementary Notes #2), despite slightly higher percentage of

139 target reads for CryptoCap_75k than CryptoCap_100k (64.73% vs. 55.6% baits,
140 respectively), both bait sets showed similar depth and breadth of coverage (Table 2; p-
141 values > 0.05; Figs. S2–S3).
142
143 *In Silico Simulations of Enrichment and Sequencing*
144 *In silico* simulations exhibited breadth of genome coverage of up to 99.9% and a
145 depth of coverage of 7.6X (for ~200,000 paired reads), minimal off-target hybridization
146 to more than a thousand non-*Cryptosporidium* genome sequences (including the human
147 genome), and reliable mapping fidelity to the database containing ten genome sequences
148 from ten *Cryptosporidium* species (10-CrypGS), which translated to high accuracy in
149 species assignment for each simulated *Cryptosporidium* species (Supplementary Notes
150 #3–7, Supplementary Tables S1–S6, Supplementary Figures S4–S13, and Supplementary
151 Results).
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 S7) 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 S8).

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 S9)
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 S9). 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 Results). Species assignments with *18S rRNA* were
170 ~70% correct (Fig. 1D; Supplementary Results).
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), the
175 percentage of reads mapped to *C. parvum* correlated with the initial concentration of
176 target DNA, with inputs below 0.0001 ng having very low percentages, with differences
177 between 1 ng and ≤ 0.01 ng, and 0.1 ng and ≤ 0.01 ng being statistically significant (p-
178 values ≤ 0.040; Fig. S18). The average depth of genome coverage was 4X for ≤ 0.01 ng,
179 29X for 0.1 ng, and 126X for 1 ng, with significant differences between 1 ng and all other
180 dilutions (p-values < 0.0001; Fig. 2A; normalized values Fig. S19). The breadth of
181 genome coverage was significantly different between comparisons below and above 0.01
182 ng (p-values ≤ 0.02; Fig. 2A; Supplementary Results; normalized values Fig. S19). For
183 unenriched libraries, accurate species assignment is observed only for 1 ng and 0.1 ng

184 inputs (Fig. 2B). For enrichments, *C. parvum* is accurately detected in inputs ≥ 0.001 ng
185 (Fig. 2B).
186 For iNextEra libraries (Supplementary Notes #10.2; Table S11, 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 Results, Fig. S21) for libraries
193 with ≤ 0.1 ng we recovered depths ≤ 4X whereas for those with 1 ng average depth was
194 34.8X (Fig. 2C). Breadth of genome coverage between the two highest input samples
195 (i.e., 1 ng and 0.1 ng, mean 57%–58%) differed from all other input samples (p values ≤
196 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* were high across all dilutions, averaging 90.5% (Table S11,
202 Fig. S22; Supplementary Results).
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 S10; Supplementary Results).

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 S12, Fig.
217 S30), the percentage of reads mapped increased from 0.02–0.15% (unenriched) to 0.37–
218 18.17% (enriched), with a fold change of 9.25–479X (Table S12, 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. In contrast, in enriched libraries,
222 DOCK and UKH101 are predominantly assigned to *C. hominis* and *C. cuniculus*, while
223 EC4 and UKP196 to *C. parvum* (Fig. S32 left). No reads mapped to the *gp60* database,
224 and very few mapped to the *18S rRNA* database with assignment to species not common
225 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
231 (Supplementary Notes #12.2.1, Table S13; Fig. S34), on average, 35.6% of enriched vs.
232 0.08% of unenriched retained reads mapped to the *C. parvum* genome sequence (p-value
233 < 0.0001; Fig. 3A; Table S13; Fig. S35). Libraries were enriched with *C. parvum* by an
234 average factor of 2,095X (7.20–5,097X; Table S13). The depth and breadth of genome
235 coverage were significantly higher for enriched samples, with average depth of 18X and
236 breadth of 31% vs. 0.04X and 0.02% in unenriched samples (Fig. 3B; normalized
237 parameters per 100,000 reads can be found in Supplementary Results and Fig. S36).
238 For unenriched libraries, qPCR C_T values were not a good predictor of the percentage
239 of reads mapping to 10-CrypGS ($R^2 = 0.012$, p-value = 0.36; Fig. 3C). In enriched
240 libraries the qPCR C_T values were good predictor of the 15.5% of the variation in the
241 percentage of mapping reads with a statistically significant effect showing that for every
242 increase of one qPCR C_T unit, the percentage decreased by 2.81% ($R^2 = 0.155$, p-value =
243 5.11e-5; ±18%; Fig. 3D; Table S13).
244 Species assignments for unenriched libraries resemble the complex background from
245 LOD experiments (cf. Figs. 2B and 3C). Following enrichment, most hits matched *C.*
246 *parvum* (63 samples have >50% of hits to *C. parvum*; Fig. 3E), but many samples have
247 moderate proportions of reads assigned to other species, including those with limited
248 evidence to infect humans (Fig. 3F; see Supplementary Results).
249
250 <u>iNextEra Libraries</u>
251 For the 91 clinical iNextEra libraries (Supplementary Notes #12.2.2, Tables 3 and
252 S14; Fig. S37), on average, 0.03% of unenriched reads, 15.85% of single-enriched reads,

253 and 64.98% of double-enriched reads mapped to the *C. parvum* genome sequence
254 (Supplementary Notes #12.2.2, p-values < 0.0001, Fig. 4A, Table S14, Fig. S37;
255 Supplementary Results), yielding an average enrichment factor of 1,112X from
256 unenriched to single-enriched, and an extra 9.24X from single- to double-enriched (Table
257 S14). The mean depth of genome coverage was significantly higher for double-enriched
258 samples (95.5X) than for single-enriched libraries (32.1X) and unenriched libraries
259 (0.05X) (p-values < 0.005; Fig. 4B; normalized values per 100,000 reads in
260 Supplementary Results and Fig. S39, Table S14). The mean breadth of genome coverage
261 was significantly higher for double- and single-enriched samples (48% and 54%,
262 respectively) than for unenriched samples (2.6%, p-values < 0.0001).
263 The percentage of reads mapped was negatively correlated with qPCR C_T values for
264 all library types (Figs. 4C–4E). In unenriched and double-enriched libraries, the linear
265 model fails to detect C_T values as predictors of percentage of reads mapped (p-value >
266 0.05; Figs. 4C and 4E), but in single-enriched libraries qPCR C_T has a significant but
267 small effect on percentage of reads mapping ($R^2 = 0.006$, p-value = 0.016), that is for
268 every increase of one qPCR C_T unit, the percentage decreased by 1.29% (Fig 4D).
269 Double-enrichments allowed for the retrieval of a higher percentage of target reads, even
270 at higher C_T values, than the other two library types (Figs. 4C–4E).
271 Species assignments were similar to those from NEB-iTru libraries in unenriched and
272 single-enriched libraries (cf. Figs. 3E vs. 4F, and Figs. 3F vs. Fig. 4G). Double-enriched
273 libraries compared to single- and unenriched libraries exhibit an increased proportion of
274 hits to the most frequently observed human-infecting species (*C. parvum* and *C. hominis*;
275 Fig. 4H). Enriched libraries mapped most frequently along either *C. parvum*

276 chromosomes for UKP samples or *C. hominis*-*C. cuniculus* chromosomes for UKH
277 samples (Fig. S40), aligning with the species identified in the multi-locus typing study in
278 the same samples[48].
279
280 iNextEra vs. NEB-iTru Libraries
281 iNextEra and NEB-iTru libraries performed similarly (Table S15) but with
282 noteworthy differences. In single-enriched libraries, NEB/iTru samples obtained and
283 retained more reads following quality filtering, and a higher percentage of reads mapped
284 to the *C. parvum* genome sequence than iNextEra samples. The average depth of
285 coverage is similar for both protocols (p-value = 0.7); however, the breadth of coverage
286 was significantly higher for iNextEra libraries (p-value < 0.0001; Supplementary Notes
287 #12.2.3, Table S15; Fig. S41–S45; Supplementary Results).
288
289 *Population Genomics Analyses of Clinical Samples*
290 Enrichment drastically improved SNP recovery: obtaining 94,783–127,102 SNPs for
291 double- and single-enriched vs. 15,430 for unenriched in raw VCF files (Figs. S46–S51),
292 with large variations in read depth (DP; Fig. S52; Supplementary Results). Filtered VCF
293 files contained: 76 and 77 samples, and 12,498 and 13,578 SNPs for single- and double-
294 enrichments, respectively (Figs. S53–S56). For unenriched libraries, filtering was based
295 on read depth and biallelic variants only, yielding 91 samples and 11,226 SNPs (Figs.
296 S57–S58; Supplementary Notes #13.4; Supplementary Results).
297 Population genomics analyses (Supplementary Notes #13.5), such as principal
298 component analysis (PCA), phylogenetic networks, genetic distance trees, and Structure,

299 are highly concordant in enriched datasets. Both datasets suggest hierarchical structure (K
300 = 2; Figs. 5, S59–S68), with distinct clustering of UKP (*C. parvum*) and UKH (*C.*
301 *hominis*) samples, aligning with the species assignment analyses from above and previous
302 multi-locus typing results[48] (cf. Figs. 4H and 5A–5B). Discrimination not found in
303 unenriched libraries (Fig. 5C). Minor distinctions between single- and double-enriched
304 libraries are evident in sample UKH149, which is positioned in the trees as more closely
305 related to *C. parvum* than to *C. hominis*. However, in double-enriched libraries, UKH149
306 clusters much closer to all other *C. parvum* samples than in single enrichments (Figs. 5A
307 and 5B).

308 High pairwise F_{ST} values between UKP and UKH samples were estimated in enriched
309 (0.94 in double vs. 0.98 in single enrichments), but not in unenriched libraries (0.084).
310 These values are balanced across all chromosomes, showing no locus bias (Figs. S69–
311 S71). Pairwise genetic distances within samples of *C. parvum* reveal two to three clusters
312 in enriched libraries (Fig. S72–S73; Supplementary Results). For *C. hominis*, enrichments
313 reveal that the three samples are highly distinct from each other, even more than the
314 differences shown between clusters of *C. parvum* (Figs. S72–S73). Genetic distances
315 using unenriched libraries fail to reflect any biologically meaningful pattern (Fig. S74).

316 Deconvolution analyses (Supplementary Notes #13.6.1) revealed different numbers of
317 inferred strains (inferred K) among library types (Figs. 5A–5C). Sample UKH149 was
318 detected as having two strains by both enrichment methods (Table S16; Fig. 5B), and two
319 more samples were detected as having two strains for double-enriched libraries (UKP347
320 and UKP391) with effective Ks above 1.2 and a proportion of the least dominant
321 haplotype larger than 10% (Table S16; Fig. 5C, Table S16; Supplementary Results).

322 Three samples (UKH149, UKP347, and UKP391) exhibit the most differences between
323 capture methods. When assessing allelic read depth (AD) and genotype quality (GQ),
324 UKH149 and UKP347 had statistically larger AD and GQ values in double-enrichments
325 compared to single-enrichments (p-values < 0.0001; Figs. S75–S78, Table S17–S18).
326 Sample UKP391 had similar AD (p-value = 0.64; Fig. S79), but GQ was statistically
327 larger in double than in single enrichments (p-values < 0.0001; Fig. S79, Table S19).
328 Genetic diversity, measured when ploidy was set to 2 in GATK (Supplementary
329 Notes #13.6.5) revealed that the nucleotide diversity (π) of polymorphic sites within
330 UKH samples (considering UKH149: single-enriched mean $\pi = 0.066$; double-enriched
331 mean $\pi = 0.335$; not considering UKH149: single-enriched mean $\pi = 0.005$; double-
332 enriched mean $\pi = 0.0204$) and within UKP samples (single-enriched mean $\pi = 0.0122$;
333 double-enriched mean $\pi = 0.0187$) was statistically different between double- and single-
334 enrichments (p-values < 0.0001); and also between UKH and UKP samples (p-values <
335 0.0002; Fig. S80). High levels of heterozygosity were observed in samples with high
336 values of effective K in DEploid. These values were consistent among chromosomes,
337 showing no locus bias (Figs. S81–S86). Double enrichments showed a higher mean
338 heterozygosity compared to single enrichments (0.009 vs. 0.002), while unenriched
339 libraries showed a mean heterozygosity 1 for all samples (Fig. S86).
340

341 **Discussion**

342 Obtaining sufficient *Cryptosporidium* DNA for genomic studies has been a challenge
343 in the quest to understand *Cryptosporidium* global diversity and its links to transmission,
344 infection, pathogenesis, and host response. We developed and evaluated a hybridization

345 capture system to obtain *Cryptosporidium* genomic data from limited amounts of input
346 DNA and evaluated its performance in a variety of settings.
347 Experiments based on simulations and on DNA from purified oocysts demonstrate
348 that CryptoCap_100k retains nearly the whole genome of target species with minimal
349 bias (Figs. 1, S15–S17). However, obtaining pure oocysts requires increased effort and
350 time spent per sample, and via propagation, strains may be altered [14,29,31]. Thus, we aimed
351 to eliminate the need for purifying oocysts to obtain usable genomic data.
352 When the starting materials is fecal DNA, CryptoCap_100k increases on-target reads,
353 even at high C_T values, enables the pooling of ~10 libraries per capture reaction, and bait
354 dilution increases affordability. This translates into inexpensive genomic data that can be
355 applied to an extensive library of samples, including archived or fresh samples, from
356 diverse origins such as stool, environmental sources, oocysts, or existing extracted DNA.
357 Key findings reveal that CryptoCap_100k: i) is highly specific and achieves a higher
358 percentage of on-target reads than unenriched libraries in clinical fecal samples with both
359 NEB-iTru (~437X increase) and iNextera (~500X in single-enriched and ~2,167X in
360 double-enriched) libraries, ii), recovers ~85% of the genome without deep sequencing, in
361 DNA samples from pure oocysts, iii) recovers ~50% of the genome in fecal DNA
362 samples, iv) facilitates robust species assignment and SNP calling for population
363 genomics and deconvolution of mixed-infections within samples, and v) double-
364 enrichments should be prioritized in samples with very low target DNA.
365
366 *Development and Assessment of CryptoCap_100k*

367 With a foundation similar to the previous whole-genome sequencing bait capture for
368 *C. parvum* (CryptoCap_75k)[46], CryptoCap_100k used more complete genome sequences
369 that included the 16 subtelomeric regions (i.e., *C. parvum* IOWA-ATCC[49]) and included
370 additional baits to account for variation among *Cryptosporidium* species that commonly
371 infect humans. Performance between CryptoCap_75k[46] and CryptoCap_100k is similar *in*
372 *silico* (Table 2, Figs. S1–S3); however, CryptoCap_100k allows for the pooling of ~10
373 libraries before enrichment, decreasing processing costs and time.
374 When assessing target *Cryptosporidium* species, CryptoCap_100k recovered large
375 breadths of the genome (~95% *in silico*, ~85% from pure oocysts, and ~50% from fecal
376 samples; Tables S1, S7, and S15) with low sequencing efforts. A higher breadth of
377 coverage was not achieved by using more baits or performing double enrichments. These
378 results suggest that breadth is a property dependent on target input, as none of the
379 samples with $C_T > 28$ reached a breadth ≥ 50% (Table S14). Still, CryptoCap_100k
380 enables obtaining genomic data for downstream analyses for 76%–85% of clinical
381 samples.
382 To make recommendations about how to use CryptoCap_100k best, we defined two
383 distinct limits of detection (LOD) for: i) minimum target genomic equivalents and ii)
384 minimum bait concentration. Given that a single haploid sporozoite contains ~9 fg of
385 DNA (with four sporozoites per oocyst) and is typically found in a complex community,
386 we diluted *C. parvum* DNA in a consistent background of calf DNA and determined the
387 data quality threshold. Enriched libraries exhibit a high percentage of target reads,
388 substantial breadth and depth of genome coverage, and accurate species assignment with
389 inputs as low as 0.01 ng of target DNA (~1,000 sporozoites or ~250 oocysts equivalents;

390 Figs. 2A–2B), but higher input of target DNA significantly reduces non-target reads (Fig.
391 S22). Non-target reads are likely caused by non-specific interactions of beads and/or
392 biotin with background DNA when target DNA is low[50]. These parameters improved
393 further in double-enriched iNextEra libraries with 0.001 ng of target DNA (~100
394 sporozoites or ~25 oocyst equivalents; Fig. 2D). These findings validate that
395 CryptoCap_100k can obtain genomic information at low DNA levels without purifying
396 oocysts. The LOD assay with bait dilutions suggests that there are no differences in
397 estimated parameters between bait dilutions (full amount down to a sixteenth, Figs. S23–
398 S27), offering a cost-effective strategy for large-scale and effective use.

399

400 *Validating CryptoCap_100k in Clinical Samples*

401 We successfully tested CryptoCap_100k on a small initial set of seven clinical
402 samples. Three of these had also been analyzed with CryptoCap_75k[46], which identified
403 UK101 as *C. hominis* and UK196 as *C. parvum*[46], both of which agreed with our analyses
404 (Fig. S32). CryptoCap_75k also identified sample EC4 as *C. canis* based on the *18S*
405 *rRNA* marker, but did not obtain sufficient genome-wide read coverage to assess SNP
406 diversity [46]. In our analyses, which lack a reference genome for *C. canis*, we identified
407 this sample as largely *C. parvum*, with a profile of likely mixed ancestry (Fig. S32), but
408 this was not validated with SNP data.

409 Several studies recommend sequencing only samples with C_T values < 16[30,51], and
410 although our set of 100 clinical samples had an average C_T value of 25.4, we were able to
411 recover significantly more target reads, higher averages of breadth and depth of genome

412 coverage, and better accuracy of species assignment in enriched vs. unenriched libraries
413 with both the NEB-iTru and iNextEra protocols (Fig. 3; Fig. 4; Table S13–S14). In the
414 linear regression models for C_T values and target reads, enrichments show less loss of
415 target reads at higher C_T values compared to unenriched libraries (Figs. 3C–3D, 4C–4E).
416 Double enrichments consistently resulted in significant increases of target reads and
417 depth of genome coverage (Figs. 4A–4B), as well as more accurate species assignment
418 than single-enrichments (Figs. 4F–4H).

419

420 *CryptoCap_100k Enables the Study of Population Genomics*

421 CryptoCap_100k is a reliable tool for high-confidence SNP calling of thousands of
422 variants. Enrichment methods resolved species-level (*C. parvum* and *C. hominis*)
423 clustering, which matched the multi-locus typing results[48] (Figs. 4H, 5A–5B). Beyond
424 detecting different species, we were able to assess the diversity within each species.

425 Enrichments show that *C. parvum* haplotypes are highly similar and have low
426 nucleotide diversity. Still, two clusters within the species are detected (Figs. S68 and
427 S73), which may correspond to two circulating but highly similar variants, or samples
428 with some levels of mixed infections; however, the latter was not confirmed by our
429 deconvolution analyses. Genomic studies have revealed two distinct clusters for *C.*
430 *parvum* in the UK[25]. Our samples may represent two less divergent lineages within just
431 one of these clusters, highlighting the prevalence of clonality and the dominance of some
432 lineages in the region (Figs. 5A and 5B).

433 We also identified two distinct clusters of *C. hominis* separated by the first PC axis,
434 which explains 56-57% of the variance (UKH148 + [UKH149 and UKH150]; Figs. 5A–
435 5B). Two subspecies of *C. hominis* have been detected in the UK, *C. h. hominis*, and *C.*
436 *h. aquapotentis*, with some evidence of gene flow and recombination between them in the
437 region[26]. The highly divergent *C. hominis* samples in our study may reflect the two
438 subspecies, given their comparable levels of variation.

439 Given that more than 90% of cryptosporidiosis cases in humans are caused by *C.*
440 *parvum* and *C. hominis*[52], our clinical results were representative. Because ~74% of the
441 baits in our set match to *C. parvum*, it is relevant to assess CryptoCap_100k performance
442 for other species. In our study, we demonstrate that CryptoCap_100k enables the
443 collection of genomic data for *C. hominis* in clinical samples, from which only ~3% of
444 the baits had been designed specifically, but we expected a larger set to work thanks to
445 the 96.8% percent similarity with *C. parvum* genome sequences [53] and the mismatch
446 tolerance of the baits. Our evidence suggests that CryptoCap_100k works appropriately
447 in other species: i) *in silico* analyses reveal proper hybridization of our baits to other
448 target species (Fig. S7); ii) *in vitro* tests using pure oocyst of *C. meleagridis* show proper
449 species assignment and a high percentage of target reads (Fig. 1B); and iii) the number of
450 baits for some species, such as *C. meleagridis* and *C. viatorum* included in
451 CryptoCap_100k is even higher than the number for *C. hominis*, which worked
452 successfully (Table 1).

453

454 *CryptoCap_100k Enables the Detection of Mixed Infections*

455 *Cryptosporidium* infections can contain multiple strains, subspecies, or species.
456 Current methods for detecting mixed infections include amplicon, PCR-RFLP, and
457 MLST approaches[34,54,55]. However, these methods have inherent biases due to
458 preferential amplification and can miss mixed infections that are not comprised of *C.*
459 *parvum* and *C. hominis*[34,35,54]. Regarding mixed-species infections, CryptoCap_100k was
460 able to detect these in *in silico* simulations (Fig. S13); in *in vitro* pools of *C. parvum* and
461 *C. meleagridis*, with a slight bias towards *C. parvum*; and in clinical samples (i.e., EC4,
462 UKH149, UKP381, and UKH151; Figs. 4H and S32). These results were confirmed for
463 sample UKH149 through deconvolution analyses; however, double-enrichments exhibit
464 higher proportions of *C. parvum* compared to single-enrichments. These higher
465 proportions of *C. parvum* can be caused by minor enrichment biases due to the high
466 number of baits designed from *C. parvum*[46].

467 Mixed infections with strains of the same species were detected in UKP347 and
468 UKP391 (Fig. 5A) in double-enriched but not in single-enriched libraries. These
469 discordances were associated with higher allelic depth (AD) and/or improved genotype
470 quality (GQ) in double-enriched libraries (Figs. S75–S79), which likely allowed for
471 greater resolution and ability to detect imbalances in within-host allele frequency by
472 DEploid.

473 In summary, CryptoCap_100k results in a cost reduction of approximately 98%
474 relative to deep sequencing of fecal samples. However, cost savings will vary depending
475 on the abundance of *Cryptosporidium* DNA in the sample (Supplementary Results and
476 Table S20). CryptoCap_100k can recover >80% of the genome at high depth from pure
477 oocyst samples (Figs. S13–S14) and achieve as high as 3,690X fold-change enrichment

478 in fecal samples, without the need for oocyst purification, allowing high-accuracy species
479 assignment (Table S14). The quality of the retrieved genomic data allows for thousands
480 of SNPs to be called, useful for population genomic and mixed-infection studies. Double
481 enrichment improves coverage for low-input DNA samples, but this outcome can be
482 matched with single enrichments and higher sequencing effort. Studies using the
483 CryptoCap_100k will enhance our understanding of the population genetics of human-
484 infecting *Cryptosporidium* species and their relationship to epidemiology. Generating
485 economical and efficient genome-level data enables the growth of reference resources
486 and the refinement of analyses and methods for subtyping.

487

488 **Methods**

489 *CryptoCap_100k Bait Design & Characterization*

490 The genome sequences of six *Cryptosporidium* species were chosen for bait
491 design (Table 1), along with 55 *18S rRNA* sequences and eight *gp60* sequences (Data
492 Availability). All sequences were soft-masked for simple and low-complexity repeats
493 using RepeatMasker 4.1.1 (http://www.repeatmasker.org/). We aimed to design 100,000
494 120-nucleotide baits. Baits that were ≤ 50% soft-masked for repeats were retained. These
495 parameters resulted in 435,323 raw baits. Then, VSEARCH 2.15.0[56] was used to cluster
496 baits at various levels of sequence overlap and identity (Table 1), allowing for
497 mismatches to reduce the number of baits while maintaining diversity. The genome order
498 was designed to prioritize genome sequences based on their completeness and likelihood
499 of human infection, with *C. parvum* IOWA-ATCC listed first for all clustering
500 procedures. Then, *C. hominis*, *C. cuniculus*, and *C. tyzzeri* were collapsed at 50%

501 overlap, and 90% identity relative to *C. parvum*; and *C. meleagridis* and *C. viatorum*
502 were collapsed at 45% overlap and 85% identity relative to *C. parvum*. From *C.*
503 *viatorum*, a subset of baits was randomly selected. Raw baits were obtained for the *18S*
504 *rRNA* targets and collapsed at 50% overlap and 95% identity (Table 1).

505 We characterized CryptoCap_100k (Supplementary Notes #1) by using twelve
506 genome sequences available for ten species of *Cryptosporidium* (Data Availability).
507 BWA 0.7.17[57] was used to map the CryptoCap_100k bait sequences to each genome
508 sequence. Alignment files were generated and converted using Samtools 1.10[57,58]. The
509 bamCoverage function from DeepTools 3.3.1[59] was used to produce bigwig files to
510 visualize coverage tracks calculated as the number of reads per bin in IGV-Web app
511 version 1.10.8[60]. To obtain alignment statistics, the depth option from Samtools v1.1.0
512 and the *-a* flag were used to get all sites and create a depth file[58]. The mean depth, the
513 proportion of baits mapped to each reference, and the breadth of genome coverage were
514 estimated. The breadth of coverage, as reported throughout this manuscript, was
515 determined by estimating the number of bases covered in the alignment over the
516 reference length, with a minimum coverage depth of 1X.

517 We also compared the CryptoCap_100k bait set to CryptoCap_75k[46]
518 (Supplementary Notes #2) by estimating the same parameters as above, and using
519 Welch's two-sample t-test for comparisons.

520

521 *In Silico Simulations of Enrichment with CryptoCap_100k & Sequencing*

522 To validate data quality from CryptoCap_100k before bait synthesis, a simple
523 infection scenario, error-free and bias-free, was simulated. We used twelve available

524 genome sequences from ten *Cryptosporidium* species to simulate the enrichment and
525 sequencing process with CryptoCap_100k, assuming two different ranges of insert size
526 fragments (Supplementary Notes #3). To simulate hybridization, we used the coordinates
527 of the mapped baits in relation to each reference genome, and we extended the position
528 by several bases to "capture" larger fragments than the bait length using Samtools[58]. To
529 simulate sequencing, ART 2016.06.05[61] was used to reproduce ~200,000 paired-end 150
530 bp fastq reads from the extended coordinates in each reference genome. Then, reads were
531 mapped to their respective genome sequences using BWA 0.7.17[57], and we estimated
532 depth and breadth of coverage. To validate the accuracy of species assignment, we used a
533 reference database containing ten *Cryptosporidium* species genome sequences (10-
534 CrypGS), a *gp60* database[2], and an *18S rRNA* database[52,62,63], and used BBMap 38.90[64] to
535 map simulated fastq reads to each and explore the proportion of hits to each
536 *Cryptosporidium* species.

537 To assess the sensitivity of the baits to off-target DNA, we assessed the affinity of
538 our baits to 1,008 genome sequences[65], including the human genome. Because *in vitro*
539 libraries were diluted in calf thymus DNA (see *In Vitro Test of Limits of Detection with*
540 *Different Levels of C. parvum Dilution*), we also assessed affinity to the *Bos taurus*
541 genome sequence (GCA_002263795.3). For both analyses, we used BWA 0.7.17[57] to
542 map the bait sequences, and we estimated the number of hits as well as the quality of the
543 mapping.

544 To explore the ability to discriminate different species within mixed-species
545 infection samples, we created four simulated samples by concatenating the fastq files
546 already simulated for each species above. The four *Cryptosporidium* mixed infection

547 samples contained at equal proportions: i) reads from twelve genome sequences from ten
548 species; ii) reads from ten genome sequences from ten species; iii) reads from five
549 genome sequences and species that share high percent identity; and iv) reads from three
550 genome sequences and species commonly found in human infections. Then, we mapped
551 these mixed reads to 10-CrypGS, the *gp60*, and the *18S rRNA* databases. See
552 Supplementary Notes #3–7 for details.
553
554 *In Vitro Enrichment & Validation with CryptoCap_100k Baits with Pure Oocysts' DNA*
555 Pure oocyst DNA was obtained from three *C. parvum* samples from Dr. Michael
556 Grigg (National Institutes of Health, US) and *C. meleagridis* DNA isolate TU1867 from
557 BEI Resources (BEI Resources NR2521 Manassa, VA). Metagenomic libraries were
558 prepared with the New England BioLabs Ultra FS II library kit (see *Metagenomic*
559 *Libraries*), and aliquots of these were left unenriched or enriched with one-tenth of the
560 recommended volume of CryptoCap_100k (see *Hybridization Capture Enrichments*).
561 Libraries were sequenced using an Illumina MiSeq with a PE250 kit.
562
563 *In Vitro Simulation of Mixed Infections*
564 To test the ability to detect mixed infection samples, after NEB/iTru library
565 preparation (see *In Vitro Enrichment & Validation with CryptoCap_100k Baits with Pure*
566 *Oocysts' DNA*) but before bait capture, two indexed libraries were pooled at equimolar
567 ratios to simulate a mixed infection. After pooling, an aliquot was enriched with one-
568 tenth of the recommended volume of CryptoCap_100k, and another remained
569 unenriched. These were sequenced in an Illumina MiSeq using a PE250 kit. We

570 estimated the number of reads per sample after demultiplexing, expecting roughly 50%
571 for each species.
572
573 *In Vitro Test of Limits of Detection with Different Levels of C. parvum Dilution*
574 To determine the CryptoCap_100k limits of detection (LOD) of *Cryptosporidium*,
575 we performed two types of library preparations, NEB-iTru and iNextEra[66], on samples
576 with known and varied input DNA. For NEB-iTru libraries, we diluted *C. parvum*
577 genomic DNA in a background of calf thymus DNA (Sigma-Aldrich, St. Louis, MO). A
578 45-ng aliquot of calf thymus DNA was pipetted into five tubes. In the first tube, 5 ng of
579 *C. parvum* genomic DNA was added to a final volume of 50 µL. Then, three 1:10 serial
580 dilutions were performed. The fifth tube corresponded to a negative control with just the
581 calf DNA. Metagenomic libraries using the NEB-iTru protocol and 10 µL of each
582 dilution were generated (see *Metagenomic Libraries*). In this experiment, up to 7 libraries
583 were pooled, an aliquot was left unenriched, and independent aliquots were enriched with
584 full, one-half, a quarter, an eighth, or a sixteenth of the recommended volume of
585 CryptoCap_100k (see *Hybridization Capture Enrichments*). Pools were sequenced at
586 Novogene using a HiSeq X with a PE150 kit. The number of reads, percentage of reads
587 mapped to the *C. parvum* genome sequence, and depth and breadth of genome coverage
588 were estimated.
589 For iNextEra libraries, 90 ng aliquots of calf DNA were pipetted into six tubes. In
590 the first tube, 10 ng of *C. parvum* genomic DNA was added in a final volume of 100 µL.
591 From this, four 1:10 serial dilutions were performed. The sixth tube corresponded to the
592 negative control (calf DNA only). iNextEra library preparation was performed with 10

593 µL inputs from these dilutions (see *Metagenomic Libraries*). Up to 6 libraries were
594 pooled, and aliquots were left unenriched or enriched with one-quarter or one-eighth of
595 the recommended baits. Aliquots of single-enriched libraries were double-enriched with
596 one-quarter of the recommended baits (see *Hybridization Capture Enrichments*). Pools of
597 unenriched, single-enriched, and double-enriched libraries were sequenced at Novogene
598 using a NovaSeq with a PE150 kit. The number of reads, the percentage of reads mapped
599 to the *C. parvum* genome sequence, the depth and breadth of genome coverage, and
600 species assignment based on the mapping of reads against 10-CrypGS were analyzed
601 among dilutions.
602
603 *In Vitro Test of Limits of Detection with Different Levels of Bait Dilution*
604 To determine the sensitivity or limits of detection (LOD) when using a lower
605 concentration of baits than recommended, we modulated the quantity of baits added to
606 the capture reaction in one experiment with NEB-iTru libraries and one experiment with
607 iNextEra libraries.
608 For both library types, the bait dilution sensitivity assay considered pools of the
609 metagenomic libraries from the *In Vitro Test of Limits of Detection with Different Levels*
610 *of C. parvum Dilution* experiment. For NEB-iTru libraries, we tested five different bait
611 quantities: full volume (5.5 µL, per the manufacturer's specifications), and dilutions in
612 water at one-half, a quarter, an eighth, and a sixteenth. For iNextEra libraries, we tested a
613 quarter and an eighth bait dilutions. Sequencing was performed as described in the
614 section above. The number of reads, the percentage of reads mapped to the *C. parvum*

615 genome sequence, and the depth and breadth of genome coverage were compared among
616 bait dilutions.
617
618 *In Vitro Enrichment & Validation with CryptoCap_100k Baits with Clinical Samples*
619 Two clinical datasets were used for real-world validation. First, Dr. Michael
620 Grigg (National Institutes of Health, US) and Prof. Rachel Chalmers's laboratory (Public
621 Health Wales—PHW— Microbiology and Health Protection, UK) provided DNA
622 extracted (Qiagen PowerSoil Kit) from seven de-identified patient fecal samples. These
623 samples were obtained under an NIAID IRB-approved protocol NCT00006150. The
624 NEB-iTru library protocol was used, pools made, aliquots were either left unenriched or
625 enriched with one-tenth of the recommended volume for baits, and libraries were
626 sequenced in an Illumina HiSeq using a PE150 kit.
627 Second, 100 DNA extractions (Qiagen QIAamp Fast DNA Stool kit) from clinical
628 samples from Prof. Rachel Chalmers's laboratory (PHW) were used. This work was
629 carried out in accordance with a material transfer agreement between UGA and PHW,
630 which did not require ethical approval. These samples had previously been screened
631 using a multi-locus typing system [48] to detect the species present, and these previous
632 results are reflected in the sample names used here (*C. parvum* = UKP and *C. hominis* =
633 UKH). The laboratory workflow used for these clinical samples in this study is depicted
634 in Fig. 6. A real-time qPCR assay was used to quantify *Cryptosporidium*, using primers
635 for the *18S rRNA* gene and slightly modified probes [67] as specified in Supplementary
636 Notes #12.2. Two types of genomic libraries were prepared for each sample (NEB-iTru
637 and iNextEra; see *Metagenomic Libraries*). Nine DNA samples were depleted when used

638 for iTru libraries, thus not used for iNextEra libraries. Aliquots of pools of ~12 libraries
639 were left unenriched or enriched. iTru-NEB libraries were enriched with one-tenth and
640 iNextEra libraries were enriched with one-eighth of the recommended volume of
641 CryptoCap_100k. For iNextEra libraries, we performed additional double enrichments on
642 single-enriched aliquots using one-quarter of the recommended volume of baits (see
643 *Hybridization Capture Enrichments with CryptoCap_100k*). All libraries were sequenced
644 using an Illumina NovaSeq X with a PE150 kit (Fig. 6).
645
646 *Metagenomic Libraries*
647 For iTru-NEB libraries, the New England BioLabs (NEB) Ultra FS II library kit
648 (New England BioLabs, Ipswich, MA) protocol was used according to the manufacturer,
649 with enzymatic fragmentation lasting 5–11 minutes, based on the integrity of DNA in gel
650 electrophoresis. A 12- or 14-cycle PCR with iTru primers[66] was used to add unique dual-
651 indexes to each ligation product. Products were cleaned with 1.25X SpeedBeads[68] and
652 quantified with a Qubit 4.0 Fluorometer (Thermofisher, Waltham, MA).
653 iNextEra library preparation was used to decrease the time, cost, and amount of
654 input DNA needed. In brief, 2.5 µL of 10 ng/µL (25 ng) of sample DNA was used for a
655 tagmentation reaction with 3 µL of 2X TMP buffer and 0.5 µL of Transposome (BLT;
656 CAT # 20015880, Illumina Inc.), incubated at 53 °C for 30 minutes. Then, PCR reactions
657 contained 9.5 µL of Q5 Buffer, 1.25 µL of 10 mM dNTPs, 0.5 µL of Q5 High Fidelity
658 Polymerase (New England Biolabs M0491L), 40.25 µL of nuclease-free water, 2.5 µL of
659 iNextEra5 indexed primer at 5 µM, 2.5 µL of iNextEra7 indexed primer at 5 µM[66], and 6
660 µL of tagmented product. Cycling conditions were: 3 minutes at 72 °C, 3 minutes at 98

661 °C, followed by ten cycles of 98 °C for 45 seconds, 62 °C for 30 seconds, and 72 °C for 2
662 minutes, with a final extension at 72 °C for 1 minute.
663
664 *Hybridization Capture Enrichments with CryptoCap_100k*
665 Metagenomic libraries were pooled (e.g., Fig. 6) in preparation for enrichment.
666 For different experiments, the input concentration per pool of samples varied (i.e., for
667 clinical samples using NEB/ iTru, the input was between 25–80 ng/µL; for clinical
668 samples using iNextEra, it was 103 ng/µL; and for the LOD experiments, it was 26–59
669 ng/µL). The enrichment process per pool followed manufacturers' standard protocols
670 (mybaits v5, Daicel Arbor Biosciences, Ann Arbor, MI), with wash and hybridization
671 temperatures at 63 °C and modifications of bait quantities as described above. After
672 enrichments, P5/P7 PCRs with 16–22 cycles were performed, followed by 1–1.25X
673 SpeedBeads clean-ups (Sigma Aldrich, St. Louis, MO).
674 For the 100-clinical sample set and the LOD experiments, double enrichments
675 were performed on the iNextEra single-captured pools (e.g., Fig. 6). For these, aliquots of
676 the cleaned single-captured pools were pooled and used independently as input for a
677 second capture using one-quarter of the recommended baits. For post-double-enrichments
678 P5/P7 PCRs with 12–16 cycles were performed, followed by 1X SpeedBeads clean-up.
679 All PCRs were performed using the KAPA HiFi Kit (Roche, Basel, Switzerland). All
680 P5/P7 PCRs consisted of 98°C for 2 minutes.; then, cycles of: 98°C for 20 sec., 60°C for
681 30 sec., 72°C for 1 min.; followed by 72°C for 5 min.
682
683 *In Vitro Bioinformatic Processing and Analyses*

684 Bioinformatic and data analysis code are available in the Supplementary Notes.
685 Sequencing data were processed using Trimmomatic 0.39[69] , and retained reads were
686 mapped to specific genome sequences or databases using BBMap 38.9[64]. Samtools[58] was
687 used to sort bam files. We determined mapping statistics, including genome depth and
688 breadth of coverage. The results of the mapping to 10-CrypGS were used to assess the
689 proportion of hits to the different *Cryptosporidium* species and produce species
690 assignment plots. For iNextEra libraries from clinical samples, we also generated bigwig
691 files, which were visualized using the IGV-Web app version 2.2.7.
692 Cleaned reads from pure oocysts libraries and *in vitro* mixed-infection libraries
693 were mapped to the corresponding target references (i.e., *C. parvum* samples to *C.*
694 *parvum* IOWA-ATCC, *C. meleagridis* samples to chromosome-level assembly[70], also to
695 10-CrypGS, to the *gp60* database, and the *18S rRNA* database—used in simulations). The
696 sequencing reads from the LOD dilution experiments (i.e., input DNA and bait dilutions)
697 were mapped to *C. parvum* IOWA-ATCC and 10-CrypGS. iNextEra reads from LOD
698 were also mapped to *Bos taurus* (GCA_002263795.3). The cleaned sequence reads from
699 clinical samples were mapped to *C. parvum* IOWA-ATCC and 10-CrypGS. For the set of
700 100 clinical samples, qPCR C_T values were used in the *parsnip 1.1.1*[71] to define linear
701 regression models predicting the percentage of reads mapped against 10-CrypGS
702 (Supplementary Notes #12).
703
704 *Population Genomics and Mixed Infection Analyses of Clinical Samples*
705 To examine the genetic variation in natural populations of *Cryptosporidium* from the
706 set of 91 clinical samples prepared using the iNextEra protocol, we performed genotype

707 calling (Fig. S87). In short, reads from each treatment (unenriched, single-enriched, and
708 double-enriched) were mapped to the *C. parvum* IOWA-ATCC genome sequence using
709 BWA 0.7.17[57] , and the results were used in GATK 4.6.0.0[72]. Duplicate reads and
710 haplotypes were determined using Haplotypecaller with GVCF mode and merged into
711 databases for each library type to perform joint genotyping using GenotypeGVCFs.
712 Importantly, HaplotypeCaller was run with --sample-ploidy set to 1. A first filtering
713 approach was performed to obtain SNPs with QD < 2, FS > 60, MQ <40, MQRankSum <
714 -12.5, and ReadPosRankSum < -8.0.

715 The resulting VCFs were explored for missing data and coverage depth per sample
716 and chromosome using vcfR 1.15.5[73]. Then, filtering using dartR 2.9.7[74] considered only
717 SNPs with coverage (DP) larger than 5 and lower than 200, that were biallelic in the
718 dataset, and with less than 40% of missing data. Individuals with > 80% missing data
719 were removed.

720 To explore the diversity and divergence among samples, filtered VCF files for each
721 treatment were used for population genomic analyses. These included Principal
722 Components Analyses performed with SNPRelate 1.40.0[75] and ade4 1.7-23[76]. We used
723 phangorn 2.12.1[77]and SplitsTree 6.4.11[78] to compute pairwise genetic distances and
724 visualize phylogenetic networks. We estimated Nei's genetic distances between samples
725 and visualized them in dendrograms with bootstrap support using poppr 2.9.6[79] and
726 plotted these with ape 5.8-1[80]. We tested K = 1–5 clusters, each with three replicates, in
727 Structure 2.3.4[81], run through structure_threader 1.3.10[82], which utilizes Structure
728 Harvester 0.6.94[83]. We used 10,000 chains for burn-in, 100,000 chains for MCMC, and
729 set ploidy to 1, with other options left as default. We used readr 2.1.5[84] and starmie

730 0.1.2[85] to graph Evanno's results and to plot Structure assignment plots. We estimated
731 pairwise F_{ST} statistics between UKP and UKH samples using hierfstat 0.5-11[86] and
732 mapped these values along *C. parvum* chromosomes using qqman 0.1.9[87]. We compared
733 Allelic depth (AD) and genotype quality (GQ) values between VCF files of the different
734 treatments and used ggridges 0.5.6[88] to plot the frequency of values across samples.[90]
735 To explore the diversity within each clinical sample, we deconvoluted multiple
736 infecting strains using DEploid 0.7.1[89]with no reference panel. The effective and inferred
737 number of strains (K) was plotted. Additionally, using filtered VCFs from the GATK
738 4.6.0.0[72] run with the setting --sample-ploidy set to 2, we estimated nucleotide diversity
739 (π) of SNPs in VCFTools 0.1.16[90] for double and single enrichments that were subset to
740 contain only UKH (*C. hominis*) or UKP (*C. parvum*) samples; and, we assessed and
741 plotted heterozygosity levels per sample. Data analyses were performed in R Studio
742 2023.06.0. Least-square comparisons of the means between experimental groups (e.g.,
743 unenriched and enriched, single-capture vs. double-capture, etc.) were performed to
744 calculate p-values using a confidence interval of 0.95 with the package: *lsmeans*[91], which
745 provides a t-test output when comparing two variables, or performs a Tukey method
746 adjustment when comparing a family of three or more estimates. Other packages used
747 were *tidyverse 2.0.0*[92], *ggplot2 3.5.2*[93], and *plotly 4.10.4*[94].
748

749 **Acknowledgments**

750 T.G. and J.C.K. disclose support from the National Institutes of Health under
751 award number NIH R01AI148667. M.G. discloses support from the Intramural Research
752 Program of the National Institute of Allergy and Infectious Diseases (NIAID) at the

753 National Institutes of Health. We thank the authors of CryptoCap_75k[46] for sharing
754 unpublished results that facilitated the work and comparisons presented here. We also
755 thank Prisha Sharma and Fiifi Agyabeng-Dadzie for sample processing, Christopher
756 Fitzgerald for sample tracking and processing, and Mustafa Nural for sample database
757 help. This study was supported in part by resources and technical expertise from the
758 Georgia Advanced Computing Resource Center, a partnership between the University of
759 Georgia's Office of the Vice President for Research and Office of the Vice President for
760 Information Technology.
761

762 **Author Contributions**

763 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
764 to conceptualization, methodology, investigation, and writing—review and editing.
765 M.S.B., R.P.B., and B.B. designed and oversaw the production and QC of the baits.
766 M.S.B., K.N.P., M.I.U.B., contributed to investigation and writing—review and editing.
767 N.J.B-V. and A.H.S. performed validation, data curation, formal analyses, and with
768 T.C.G. writing—original draft. G.R., R.M.C., E. A-F, M.E.G., and J.C.K. contributed
769 samples and analyses to the manuscript. N.J.B-V, M.S.B, A.K., R.P.B., M.E.G., J.C.K.,
770 and T.C.G contributed to funding acquisition. All authors read and approved the final
771 version of the manuscript.
772

773 **Competing Interests**

774 B.B. and M.B. are employed by Diacel Arbor Biosciences, which manufactures and
775 sells the baits described herein. The CryptoCap_100k bait set is available for commercial

776 purchase (Design ID: D10006Crypto, Cat# #308508.V5, #308548.V5, and #308596.V5,
777 Arbor Biosciences, MI, USA).
778
779 **Data Availability**
780 The six genome sequences that supported the CryptoCap_100k bait design were
781 from GenBank: GCA_015245375.1, GCA_004337795.1, GCA_001483515.1,
782 GCA_007210665.1, GCA_004337835.1, and GCA_001593445.1. The extra sub-
783 telomeric sequences from *C. parvum* were: MZ892386.1, MZ892388.1, and
784 MZ892387.1.
785 The *18S rRNA* sequence data that supported the CryptoCap_100k bait design
786 were from GenBank with the EU250845, LC089976, DQ288166, EF641022, GU951714,
787 AF108864, LC012016, GQ121020, HQ397716, KR296812, EF641014, FJ435960,
788 JX416368, KP099082, AY120903, AY120902, JX644908, AY324641, AY324638,
789 AF159113.1, JQ413438, AF108861, GQ121021, KP730304, AY737602, AY731235,
790 JX294358, AY268584, KR819168, AY120904.1, KU608308, AF112576, EU553551,
791 JF710247, KP004205, JQ002555, KT235702, AB909498.1, AF316630, AY504515,
792 AY504516, AF262325, GQ227479, EU827297, AF093500, AF093497, AY954886,
793 HM116385, DQ650344, KY490554, KC305650, EF547155, FJ769050, HM243547, and
794 AY524773 accession codes.
795 The *gp60* sequence data that supported the CryptoCap_100k bait design were
796 from GenBank with the FJ490060, FJ490087, FJ490058, JX412915, JX412926,
797 JX412925, KC204983, and KC204984 accession codes.

798 The twelve *Cryptosporidium* genome sequences used for CryptoCap_100k bait
799 set characterization and *in silico* simulations were from GenBank with the
800 GCF_00000165345.1, GCA_004337835.1, GCA_004337795.1, GCA_007210665.1,
801 GCF_001865345.1, GCA_001865355.1, GCA_001593455.1, GCF_000006515.1,
802 GCA_000165345.1, GCA_001483515.1, and GCA_002223825.1 accession codes *C.*
803 *hominis* UdeA01 and *C. parvum* IOWA-ATCC were removed for those analyses with
804 only ten genome sequences reference database (10-CrypGS).
805 Generated fastq files from all *in silico* and *in vitro* experiments for mock and
806 clinical samples are available in GenBank BioProject number PRJNA1061798. Clinical
807 samples were scrubbed of human DNA upon submission by NCBI using their Human
808 Read Removal Tool.
809
810 **Computer Code**
811 The code and data used for all sections of this paper are described in the
812 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. Nader, J. L. *et al.* Evolutionary genomics of anthroponosis in *Cryptosporidium*. *Nat Microbiol* **4**, 826–836 (2019).

16. Baptista, R. P., Cooper, G. W. & Kissinger, J. C. Challenges for *Cryptosporidium* Population Studies. *Genes (Basel)* **12**, 894 (2021).

17. 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).

18. Šlapeta, J. Cryptosporidiosis and *Cryptosporidium* species in animals and humans: a thirty colour rainbow? *Int J Parasitol* **43**, 957–970 (2013).

19. 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).

20. 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).

21. 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).

22. 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).

23. 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).

24. Hawash, Y. DNA extraction from protozoan oocysts/cysts in feces for diagnostic PCR. *Korean J Parasitol* **52**, 263–271 (2014).

25. 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).

26. 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).

27. 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).

28. 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).

29. 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).

30. 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).

31. 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).

32. 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).

33. 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).

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

35. 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).

36. 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).

37. 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).

38. 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).

39. 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).

40. Korpe, P. S. *et al.* Case-Control Study of *Cryptosporidium* Transmission in Bangladeshi Households. *Clin Infect Dis* **68**, 1073–1079 (2019).

41. 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).

42. 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).

43. Gaudin, M. & Desnues, C. Hybrid Capture-Based Next Generation Sequencing and Its Application to Human Infectious Diseases. *Frontiers in Microbiology* **9**, (2018).

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

45. Quek, Z. B. R. & Ng, S. H. Hybrid-Capture Target Enrichment in Human Pathogens: Identification, Evolution, Biosurveillance, and Genomic Epidemiology. *Pathogens* **13**, 275 (2024).

46. 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).

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

48. 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.

49. 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).

50. 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).

51. 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).

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

53. Agyabeng-Dadzie, F., Xiao, R. & Kissinger, J. C. *Cryptosporidium* Genomics — Current Understanding, Advances, and Applications. *Curr Trop Med Rep* **11**, 92–103 (2024).

54. 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).

55. 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).

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

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

58. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

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

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

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

62. 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).

63. 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).

64. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. (2014).

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

66. 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).

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

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

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

70. 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.

71. Kuhn, M. & Davis, V. parsnip: A Common API to Modeling and Analysis Functions. (2025).

72. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

73. 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).

74. 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).

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

76. Dray, S. & Dufour, A.-B. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software* **22**, 1–20 (2007).

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

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

79. 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).

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

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

82. 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).

83. 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).

84. 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).

85. Tonkin-Hill, G. & Lee, S. Starmie: Population structure model inference and visualisation. (2016).

86. Goudet, J. hierfstat, a package for r to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**, 184–186 (2005).

87. 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).

88. Wilke, C. ggridges: Ridgeline Plots in 'ggplot2'. (2024).

89. 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).

90. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

91. Lenth, R. V. Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software* **69**, 1–33 (2016).

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

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

94. 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. Bait characterization for CryptoCap_100k and CryptoCap_75k by mapping all baits to all available reference genome sequences. Mean depth, percent breadth of genome coverage, and percentage of baits mapped to each species' genome sequence. We break down by species included and excluded from bait design.

		CryptoCap_100k 100,000 baits (This study)			CryptoCap_75k 68,966 baits (Khan and Alves-Ferreira et al., 2024)		
Species Information		Mean depth	% Breadth of coverage	% Baits mapped	Mean depth	% Breadth of coverage	% Baits mapped
Species included in CryptoCap_100k bait design	*C. parvum* IOWA II	1.086	97.55	84.98	0.909	90.89	100
	C. parvum IOWA ATCC	1.089	97.83	85.41	0.907	90.63	99.96
	C. hominis UdeA01	1.095	97.75	85.36	0.904	90.33	99.31
	C. hominis 30976	1.097	97.88	85.64	0.903	90.25	99.39
	C. cuniculus UKCU2	0.853	77.86	70.09	0.751	74.93	86.50
	C. tyzzeri UGA55	1.074	96.75	83.66	0.900	89.95	98.75
	C. meleagridis UKMEL1	0.997	91.42	80.45	0.793	79.23	91.27
	C. viatorum (partial inclusion) ABER_CVIA_1.0	0.714	66.68	60.99	0.529	51.40	64.36
Species not included in CryptoCap_100k bait design	*C. andersoni* 30847	0.008	0.52	1.08	0.005	0.45	0.87
	C. baileyi TAMU-09Q1	0.030	2.84	3.45	0.025	2.43	3.70
	C. muris RN66	0.008	0.60	1.12	0.005	0.47	0.91
	C. ubiquitum 39726	0.262	25.39	24.81	0.239	23.86	31.69

Table 3. Summary statistics from enriching 100 human fecal DNA samples with two library preparation methods using CryptoCap_100k baits and obtaining Illumina PE150 reads. Source tables are Tables S13 and S14.

	iTru Libraries			iNextEra Libraries		
Average values	Not enriched	Single enriched		Not enriched	Single enriched	Double enriched
% retained after quality filtering	61.53	83.87		89.70	90.79	94.48
% reads on target *C. parvum*	0.04	35.62		0.033	15.85	64.98
% reads on target 10-CrypGS	0.11	41.23		0.100	17.90	65.78
Enrichment factor *	N/A	2,095		N/A	1,112	10,700**
Depth per 100k *	0.00	1.12		0.001	0.563	2.702
Breadth per 100k *	0.00	2.25		0.065	1.79	5.99
Number of samples	73	100		91	91	91

*(matching *C. parvum*)

** This value corresponds to the enrichment factor of double-enriched samples compared to non-enriched counterparts, which is not displayed in Tables S13 and S14.

Figures

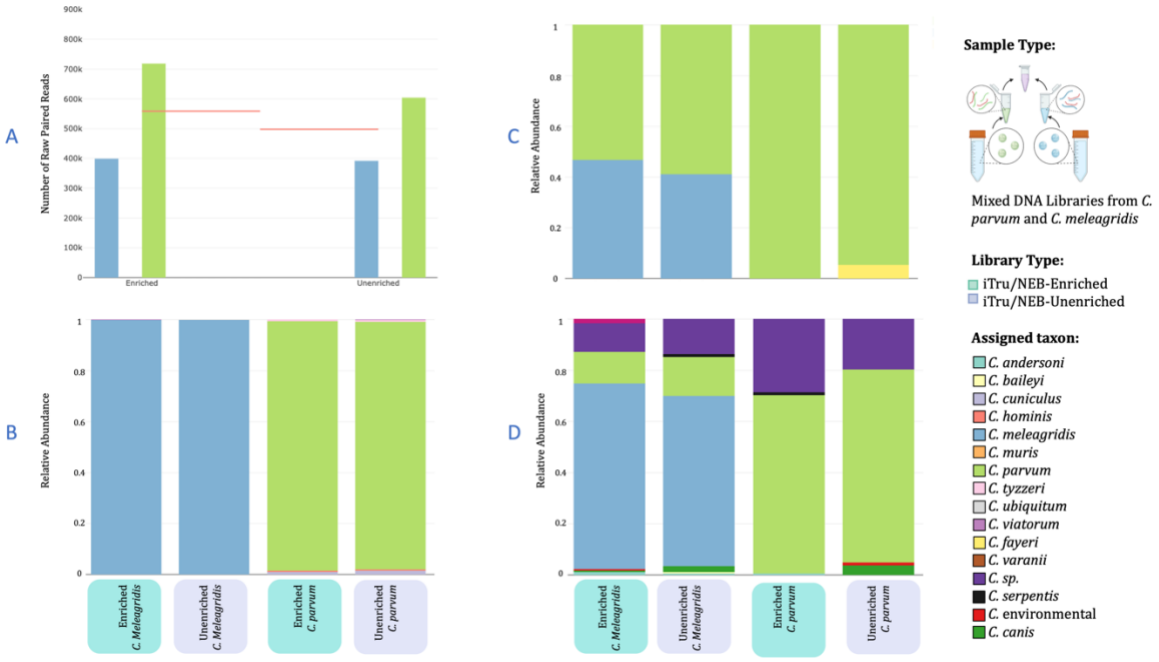

Figure 1. Species assignment plots based on mapping reads from unenriched and enriched libraries for one mock mixed-infection sample. (A) The number of raw paired reads obtained for each species in the mix, *C. parvum* in green and *C. meleagridis* in blue, for both library types. The pink lines indicate 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.

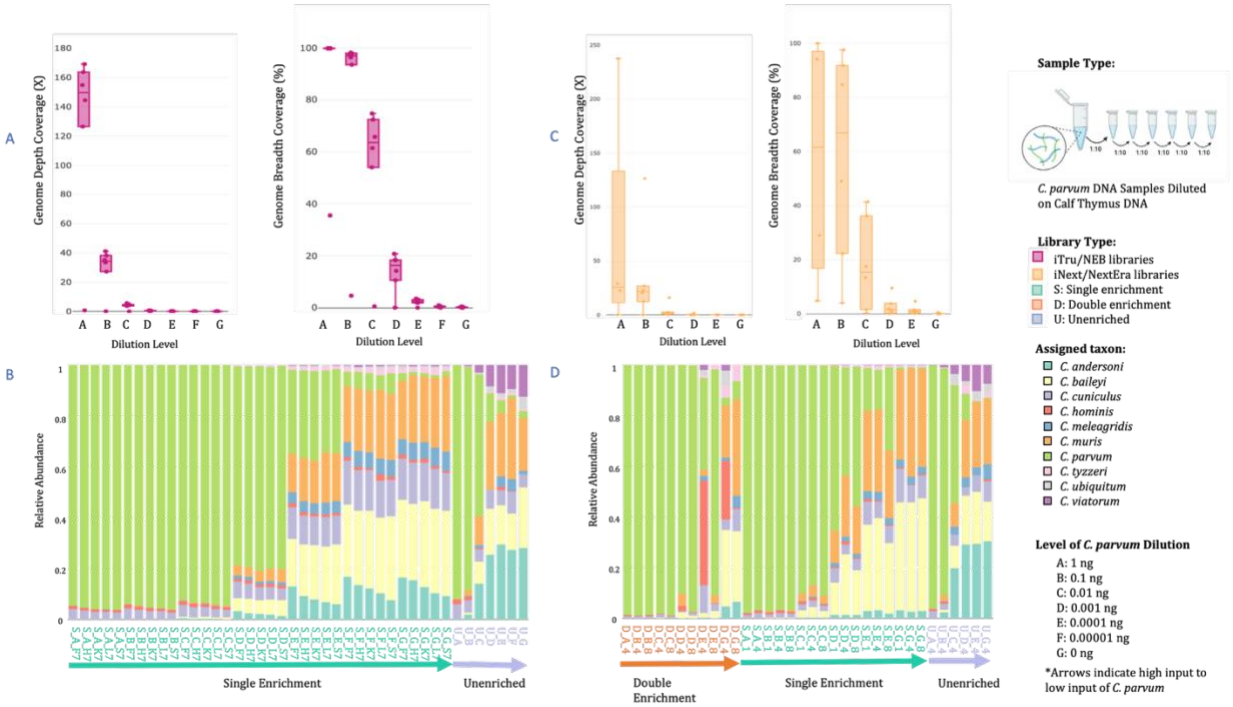

Figure 2. Summary statistics for sequence data from serially diluted *C. parvum* samples to assess detection levels given different input amounts. (A) *C. parvum* genome depth and breadth of coverage for diluted DNA samples prepared with NEB/ iTru libraries (B) species assignment of diluted *C. parvum* 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).

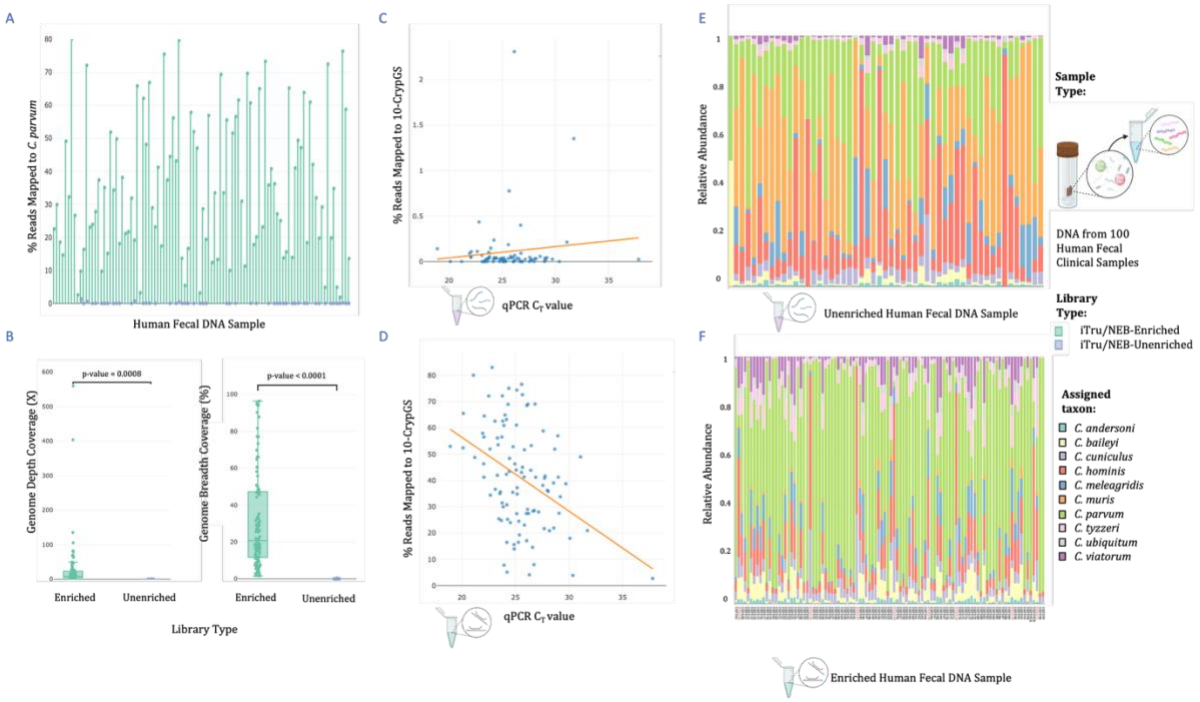

Figure 3. Summary statistics for sequence data from the 100 clinical samples prepared with the NEB protocol and iTru primers, comparing libraries enriched with CryptoCap_100k bait set versus unenriched libraries for validation. (A) Scatter plot of percentage of reads mapped from one round of enrichment compared to unenriched reads for each sample (x-axis labels not shown but in order from left to right sample 1 to sample 100). (B) Genome depth and breadth of coverage for each sample in enriched libraries compared to unenriched. 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) enriched libraries with CryptoCap_100k. Species assignment when mapped to 10-CrypGS for (E) libraries left unenriched, and (F) enriched libraries with CryptoCap_100k. Sample names for plot 3E are not shown.

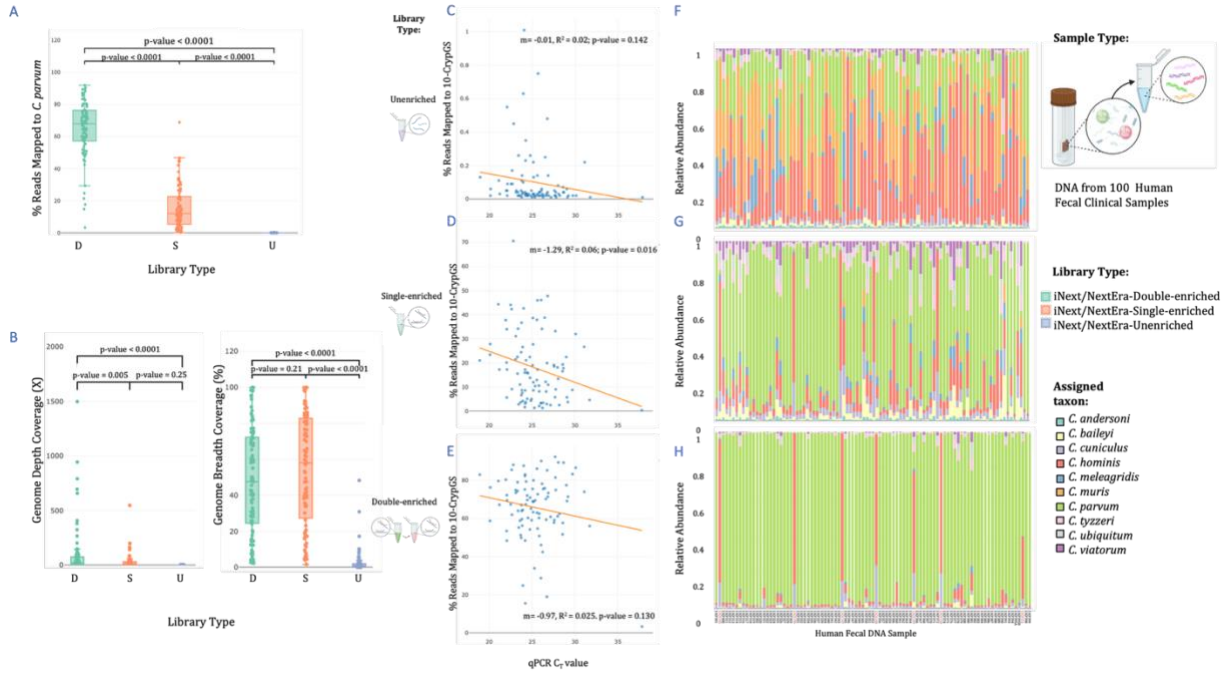

Figure 4. Summary statistics for sequence data from 91 clinical samples prepared with the iNextEra protocol and iNext primers, comparing libraries enriched with CryptoCap_100k bait set versus unenriched libraries for validation. (A) Boxplot of percentage of reads mapped to *C. parvum* from double-enriched [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.

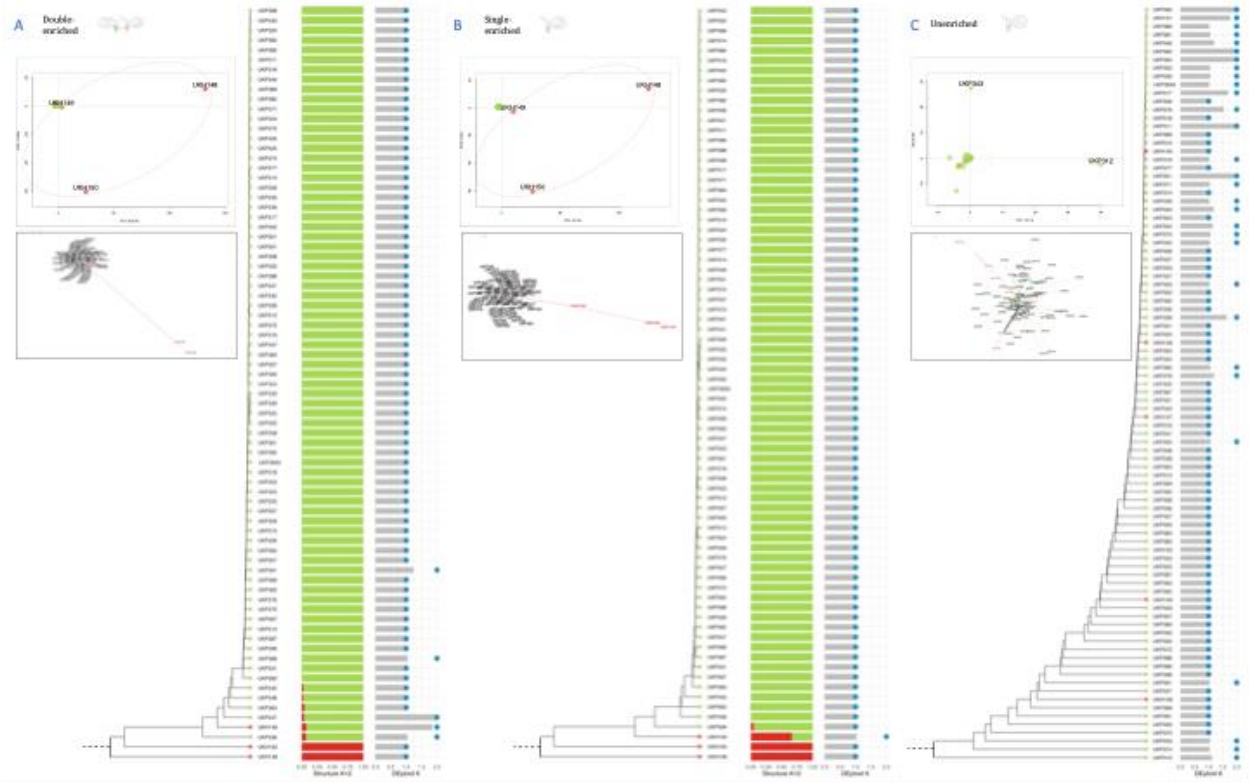

Figure 5. Population genomic analyses from called SNP variants of clinical samples prepared with the iNextEra protocol and iNext primers. Libraries were double-, single-enriched, and unenriched, respectively (A, B, and C). Each panel includes: Principal Components Analyses, phylogenetic distance networks, 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.

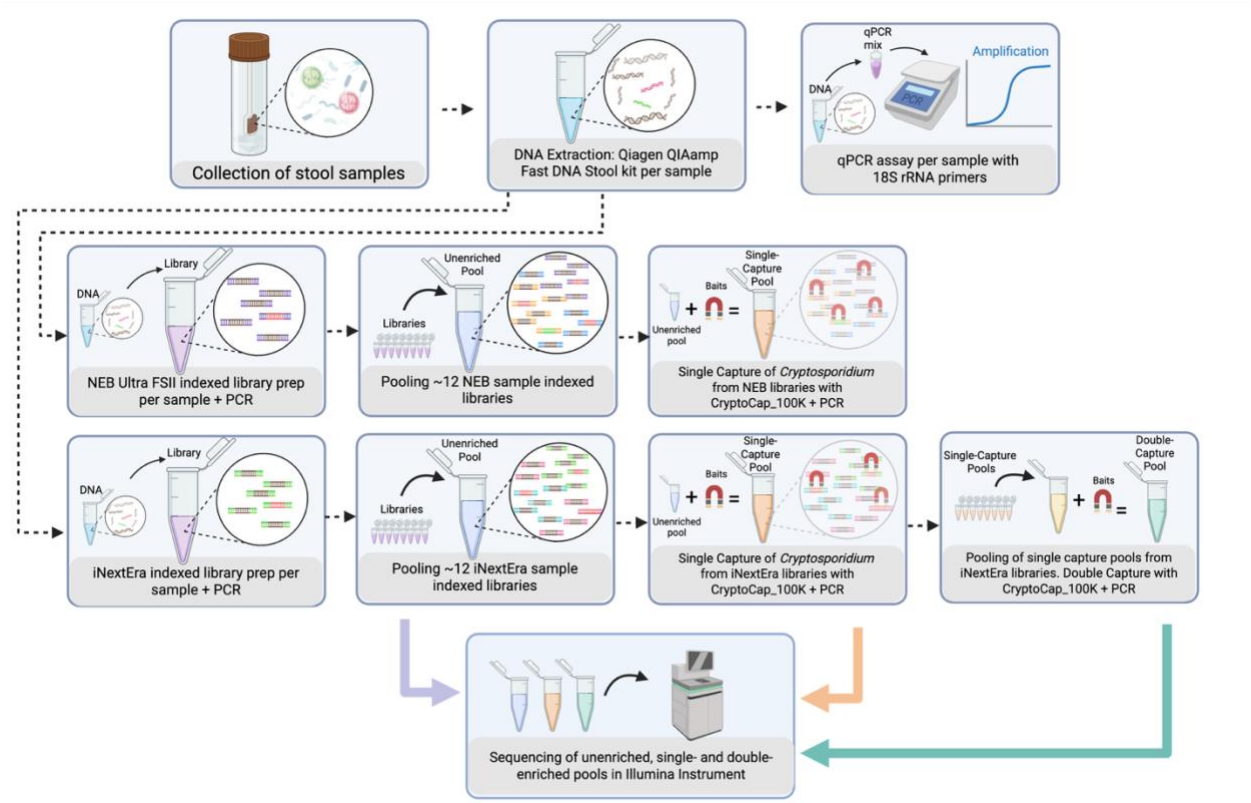

Figure 6. Laboratory workflow used for the analysis of a set of clinical samples from the UK. The genomic DNA of each patient's sample was extracted, and *Cryptosporidium* was validated via qPCR. Each DNA extract was used for two types of library prep, NEB Ultra FSII and iNextEra library prep. Unenriched 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.