WHOLE 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], Petersen KN[7], Bhuiyan MIU[2], Brunelle B[4], Robinson G[8,9], Chalmers RM[8,9], Alves-Ferreira EVC[10], Grigg ME[10], AlvesFerreira Kissinger JC[3,11,12], Glenn TC[2,3,11].

[1]Department of Biology, Oxford College of Emory University, 801 Emory Street, Oxford, Georgia, 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, USA

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

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

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

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

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

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

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

[12]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 protozoan parasites that cause severe illness in vulnerable human populations. Obtaining pure *Cryptosporidium* DNA from clinical and environmental samples is challenging because the oocysts shed in contaminated feces are limited in quantity, difficult to purify efficiently, may derive from multiple species, and yield limited DNA (<40 fg/oocyst). Here, we develop and validate a set of 100,000 RNA baits (CryptoCap_100k) based on six human-infecting *Cryptosporidium* spp. (*C. cuniculus*, *C. hominis*, *C. meleagridis*, *C. parvum*, *C. tyzzeri*, and *C. viatorum*) to enrich *Cryptosporidium* spp. DNA from a wide array of samples. We demonstrate that CryptoCap_100k increases the percentage of reads mapping to target *Cryptosporidium* references in a wide variety of scenarios, increasing the depth and breadth of genome coverage, facilitating increased accuracy of detecting and analyzing species within a given sample, while simultaneously decreasing costs, thereby opening new opportunities to understand the complex biology of these important pathogens.

45 **Introduction**

46 Cryptosporidiosis, a disease characterized by mild to severe gastrointestinal
47 symptoms, is caused by protist parasites in the genus *Cryptosporidium*. Infection can
48 damage the intestinal epithelium, leading to adverse health outcomes in vulnerable
49 human populations (young, old, immunocompromised, and pregnant)[1–3]. Transmission of
50 *Cryptosporidium* spp. to humans may occur through direct or indirect transmission[1,4–7].
51 An estimated 7.5M global cases of cryptosporidiosis occur in children aged 0-24 months,
52 and 133,422 deaths occur annually[8–10]. In the United States, *Cryptosporidium* is the third
53 leading cause of zoonotic gastroenteritis and the leading cause of waterborne disease
54 outbreaks[11]. In low-middle income countries, *Cryptosporidium* spp. is the second most
55 common pathogen in infants from 0-11 months[12]. In children under five years, infection
56 with *Cryptosporidium* spp. is the fourth leading cause of diarrheal death globally[3,12,13].
57 Furthermore, cryptosporidiosis is one of the original AIDS-defining illnesses and remains
58 associated with high mortality in immunocompromised individuals[14–16]. Thus,
59 cryptosporidiosis is globally important, affecting millions annually.

60 There are at least 38 recognized species within the genus *Cryptosporidium*[17] with a
61 wide range of host specificity and likelihood of infecting humans. Thus, *Cryptosporidium*
62 spp. vary in public health importance[17]. At the genome sequence level, based on available
63 sequences, there is a cluster of species (called the high-similarity group) that share greater
64 than 80% identity across much of their genome[5,18–20], which includes *C. cuniculus*, *C.*
65 *hominis*, *C. meleagridis*, *C. parvum*, *C. tyzzeri,* and *C. viatorum*. Host-derived and
66 environmental samples can also contain more than one species or genotype of
67 *Cryptosporidium* spp.[6,21–25]. Mixed infections may be limited to specific geographic

68 areas, particularly in endemic regions[26]. Most reported mixed infections contain *C.*
69 *parvum* and *C. hominis*[21,23,27]. Thus, methods to detect and characterize multiple species
70 and mixed infections are needed to understand cryptosporidiosis better.

71 Several molecular methods have been developed to gain insight into the genetic
72 diversity of *Cryptosporidium* species, often relying on single marker typing of conserved
73 (small subunit rRNA, SSU) or polymorphic regions (glycoprotein gene 60, *gp60*, or heat
74 shock protein 70, *hsp70*)[15,28]. Other approaches for genetic characterization include multi-
75 locus sequence typing (MLST)[29–32], and more recently, whole genome sequencing. The
76 disadvantages of MLST include the possible unreliability of markers in different regions
77 due to the limited information about the genetic variability of populations within
78 *Cryptosporidium* species[32], and thus, assays are not universal. Challenges with
79 comprehensive genomic assays such as whole genome sequencing are constrained by
80 difficulty in purifying oocysts[33] and the limited amount of DNA from purified samples
81 and from fecal clinical samples (which also have an abundance of other organisms'
82 DNA). These challenges have translated into the need for laborious purification methods
83 combined with costly deep sequencing[34,35] and the risk of failing to detect low-abundance
84 species within mixed infection samples[28]. Therefore, methods that can characterize
85 *Cryptosporidium* genomes from limited amounts of DNA from pure oocysts or fecal
86 samples that can be implemented easily and reliably would be beneficial.

87 Hybridization sequence capture (a.k.a., capture enrichment sequencing) has well-
88 known advantages[36–38] relevant to studying *Cryptosporidium* spp. and other parasites: it is
89 feasible to design thousands of baits to capture entire genomes (e.g., 9 Mb genome of
90 *Cryptosporidium* spp.); baits can tolerate many mismatches (thus species with < 20%

91 sequence divergence may be captured); baits can account for varying levels of sequence
92 divergence (thus conserved regions can anchor comparisons of related taxa); and
93 successful enrichment reduces the amount of sequencing needed (thereby reducing
94 sequencing costs). These strengths are particularly important when samples are highly
95 variable or poorly understood. The first *Cryptosporidium* whole genome bait set
96 (CryptoCap_73k; Khan and Alves-Ferreira et al., unpublished) was designed using the *C.*
97 *parvum* IOWAII incomplete genome assembly[39], and is reported to perform well when
98 each sample is enriched individually. However, library pools that contain diverse species
99 showed bias during enrichment for *C. parvum* over other distantly related species (Khan
100 and Alves-Ferreira et al., unpublished). Because fecal samples can contain unknown or
101 mixed *Cryptosporidium* species, a second-generation bait set is needed to enrich genomes
102 for a broader spectrum of *Cryptosporidium* species and consider highly polymorphic
103 genomic regions. Enriching pools of sample libraries is also desired for time- and cost-
104 efficiency.

105 To ameliorate the current challenges in the genomic study of human-infecting
106 *Cryptosporidium* spp., we developed, validated, and critically assessed CryptoCap_100k
107 and compared its performance relative to current molecular methods (i.e., *gp60*, *18S*
108 *rRNA*, and CryptoCap_73k) for the accurate identification of *Cryptosporidium* species.
109 CryptoCap_100k was designed from the most complete publicly available genome
110 sequences (Table 1). We validated the sensitivity and specificity of CryptoCap_100k *in*
111 *silico* and *in vitro* with simulated and mock samples, multiple bait dilutions, and library
112 preparation methods. Then, we tested the reliability of these methods with real clinical
113 samples. We demonstrate that CryptoCap_100k is an efficient and economical method

114 that aids in characterizing *Cryptosporidium* genome(s) in DNA extracted from pure or
115 fecal samples. The increased representation facilitates intra-sample genetic variation
116 analyses and improves species identification accuracy, opening new opportunities to
117 understand the complex biology of these important pathogens.
118
119 **Results**
120 *CryptoCap_100k Bait Design & Characteristics*
121 We designed a set of 100,000 120-mer RNA baits targeting six *Cryptosporidium*
122 species (Table 1). The *C. parvum* reference sequence required 73,672 baits for full
123 coverage. An additional 4,447 baits were needed to cover the most divergent regions of
124 the three most similar genome sequences (*C. hominis, C. cuniculus, and C. tyzzeri*).
125 Many more baits were needed to cover *C. meleagridis* and *C. viatorum*, and with relaxed
126 matching criteria, we included only a subset of sequences identified for *C. viatorum*.
127 Finally, we added baits for genes of particular interest (130 baits). The bait set is
128 available for commercial purchase (Design ID: D10006Crypto, Cat# #308508.V5,
129 #308548.V5, and #308596.V5, Arbor Biosciences, MI, USA).
130 To characterize this bait set, we mapped the baits to each of twelve whole-genome
131 sequences from ten *Cryptosporidium* species available to date (Table 2, Supplementary
132 Notes #1, Data Availability). Mean depth and breadth of coverage were high for the six
133 species used for bait design (1.0X and 90.46%, respectively). These measures were much
134 lower for four distantly related species not considered in bait design (0.08X and 7.33%;
135 Table 2; Fig. S1). When comparing the multi-species design (CryptoCap_100k) to the
136 single-species design (CryptoCap_73k) (Supplementary Notes #2), although the mean

137 percent of baits mapped to the twelve genomes was slightly higher for CryptoCap_73k
138 than CryptoCap_100k (64.73% vs. 55.6% baits, respectively), our bait set showed
139 slightly higher mean depth and breadth across all genomes (Table 2). However, these
140 differences were not statistically significant (p-values > 0.05) (Figs. S2-S3).
141
142 *In Silico Simulations of Enrichment and Sequencing*
143 After simulating library construction, enrichment, and sequencing with our bait set
144 for each of the twelve genome sequences and for two fragment sizes (Supplementary
145 Notes #3), we observed up to 99.9% breadth of coverage with an average depth of
146 coverage of 7.6X among all genomes (Table 3, Fig. S4). Both measures tend to increase
147 with fragment size in all species except *C. cuniculus*. Plots of breadth and depth of
148 coverage for those genomes with chromosome-level assemblies (*C. hominis* UdeA01 and
149 *C. parvum* IOWA-ATCC) corroborated the continuity of simulated reads from large
150 fragments across all chromosomes (Figs. S5- S6).
151 When simulated reads were mapped to the reference database containing ten genomes
152 from ten different *Cryptosporidium* species (10-CrypGS), at least 99.1% of reads for each
153 sample mapped to the corresponding genome, allowing high accuracy for species
154 assignment for all *Cryptosporidium* species tested, including those not considered in the
155 bait design (Supplementary Notes #4, Table S1). Differences in species assignment
156 between short and large fragments were largely unimportant, except for *C. hominis* (Fig.
157 1A and 1B). However, species assignment was impacted when mapped to specific gene
158 databases instead (Figs. 1C-F; Tables S2, Supplementary Results).

159 Only 894 of our baits (0.01%) map to 111 genome sequences from the 1,008 genomes
160 database[40], including the human genome (Supplementary Notes #5). In total, 84.2% of
161 these baits mapped to eukaryotic, 13.9% to bacteria, and 1.9% to archaea (Fig. S7; Table
162 S4). Most of the baits mapped with the lowest score (682 MAPQ == 0), and only 75 baits
163 had mapping scores above 20 against 31 non-*Cryptosporidium* organisms (Fig. S8-S9).
164 Simulated mixed *Cryptosporidium* infections (Supplementary Notes #7) showed a
165 high proportion of reads (≧99.88%) mapped to 10-CrypGS (Table S5). All species in a
166 sample were correctly identified for all datasets, with only slight deviations from
167 expected relative abundances (Fig. 2A, Table S6). In comparison, mapping mixed reads
168 to *gp60* or *18S rRNA* provided poorer proper species assignments (Figs. 2B, 2C; Tables
169 S7-S10; Supplementary Results).
170
171 *In Vitro Simulation of Pure Infections and Mixed Infections*
172 To further validate our bait set, the DNA of *C. parvum* and *C. meleagridis* were
173 independently used for library construction, enrichment, and sequencing (Supplementary
174 Notes #8; Table S11). A higher proportion of filtered reads mapped to the respective
175 genome sequences in enriched than unenriched libraries (95.8% vs. 68.1%, p-value =
176 0.094; Fig. S12), reaching a fold change between 1.11X and 1.35X. A similar breadth of
177 coverage (85.2% vs. 81.6%, p-value = 0.75) and greater depth (17.3X vs. 5X, p-value =
178 0.01) were observed between enriched and unenriched, respectively (Fig. S13-S14).
179 Mapping reads to 10-CrypGS showed that for both library types, a very low proportion of
180 *C. parvum* reads were assigned to two other species (*C. cuniculus* and *C. hominis)*; these

181 incorrect assignments were only slightly elevated in the unenriched samples (Fig. S15;
182 Table S12).
183 Then, individually indexed libraries from pure *C. parvum* and *C. meleagridis* DNA
184 samples were pooled to mock a mixed infection (Supplementary Notes #9; Table S13).
185 The reads from the unenriched mock sample showed that the mixture was unequal
186 initially (60.6% *C. parvum* and 39.4% *C. meleagridis*; Fig. 3A) and that enrichment
187 increased the bias slightly (64.3% *C. parvum* and 35.7% *C. meleagridis*; Fig. 3A; Table
188 S13).
189 Mapping reads from the mock mixed sample to the three databases performed
190 similarly to the *in silico* simulations (Fig. 3B-D; Supplementary Results). The 10-
191 CrypGS yielded 95.7% retained reads for enriched samples and 81.4% for unenriched (p-
192 value = 0.50). The percentages of reads that were properly assigned were very high (97%
193 unenriched, 98% enriched for *C. parvum*; and > 98% for unenriched and enriched for *C.*
194 *meleagridis*; Fig. 3B). The reads mapped to the *gp60* database performed well for the *C.*
195 *parvum* sample, but less than half of the reads were correctly assigned for the *C.*
196 *meleagridis* sample (Fig. 3C; Supplementary Results). Most reads mapped to the correct
197 species in the *18S rRNA* database, but percentages were substantially misassigned (Fig.
198 3D; Supplementary Results).
199
200 *LOD: In Vitro Sensitivity Test with Different Inputs of C. parvum DNA*
201 We tested different levels of input DNA with serial dilutions to determine the limit of
202 detection (LOD) of target genomic DNA using both iTru and iNextEra libraries. For iTru
203 libraries (Supplementary Notes #10.1), the percentage of reads mapped per sample to *C.*

204 *parvum* positively correlated with the initial concentration of DNA. That is, with low
205 inputs of 0.0001 ng, 0.00001 ng, and the negative control (0 ng), low percentages of reads
206 mapping were obtained. Statistically significant differences in the percentage of reads
207 mapped were found between 1 ng and values of 0.01 ng and below (p-values ≦ 0.004)
208 and 0.1 ng and 0.01 ng and below (p-values ≦ 0.040; Fig. S17). The depth of *C. parvum*
209 genome coverage per 100,000 reads was below an average of ~1X for dilutions equal to
210 or below 0.01 ng, 2.2X for 0.1 ng, and 2.5X for 1 ng, with significant differences
211 between the means of 1 ng and 0.1 ng and all other dilutions (p-values < 0.0001; Fig. 4A;
212 not normalized values Fig. S18). The breadth of genome coverage normalized per
213 number of reads did not correlate with input (Fig. 4A; Supplementary Results), with
214 significant differences between the input of 0.01 ng, which had a high percentage of
215 breadth (14.7%), and all other comparisons (Supplementary Results; not normalized
216 values Fig. S18). When mapped to the 10-CrypGS, we observed that for unenriched
217 samples, accurate detection is observed only for 1 ng and 0.1 ng input. Lower
218 concentrations show the effects of background DNA (Fig. 4B, unenriched). For single
219 enrichment libraries, we observe that *C. parvum* is detected at high proportions in inputs
220 of 0.001 ng or higher (Fig. 4B, single enrichment, samples A-D), with low inputs
221 (samples E-G) showing background effects.

222 For the dilution experiment with iNextEra libraries, we observed that the number of
223 raw reads obtained varied among library types and dilutions (Supplementary Notes #10.2;
224 Table S15, Fig. S19). Regardless of library type, we found that the average percentage of
225 reads mapped to the *C. parvum* genome increased with input, ranging from 0.03% in the
226 negative control to 55.39% with 1 ng of input (Fig. S20). The percentage of mapped

227 reads was lower for inputs of ≤ 0.01 ng vs. 1 ng (p-values ≤ 0.015) and for inputs ≤ 0.001
228 ng vs. 0.1 ng (p-values ≤ 0.05). The depth and breadth of *C. parvum* genome coverage
229 increased as input DNA increased (Fig. S21). However, only the breadth of genome
230 coverage between the two highest inputs (1 ng and 0.1 ng) differed from all other input
231 samples (p-values ≤ 0.017). When normalizing by the number of retained reads, depth
232 and breadth are significantly higher when 1 ng was compared to < 0.01 ng inputs (p-
233 values ≤ 0.034, Fig. 4C), and the comparisons of 0.1 ng with 0.001 and negative control
234 (p-values ≤ 0.046). When reads were mapped to 10-CrypGS, the correct assignment of
235 reads was highly correlated with input DNA (Fig. 4D). Accurate detection is observed in
236 unenriched samples with inputs ≥ 0.1 ng, in single-enriched samples with ≥ 0.01 ng, and
237 in double-enriched samples consistently with ≥ 0.001 ng, and less consistently in samples
238 of 0.0001 ng (Fig. 4D).
239
240 *LOD: In Vitro Sensitivity Test with Different Dilution Levels of Baits*
241 We tested a series of bait dilutions (1/2 – 1/16) to determine the minimum quantity of
242 baits that could be used in a capture reaction to achieve a similar performance as the
243 undiluted baits, using both iTru and iNextEra libraries. For iTru libraries, no difference
244 was observed in bait dilution for the number of reads obtained (Supplementary Notes
245 #11.1; p-values ≥ 0.975; Fig. S22; Table S14) or the percentage of reads mapping to *C.*
246 *parvum* (Fig. S23). The depth and breadth of genome coverage decreased slightly as the
247 bait dilution progressed; however, these differences were not statistically significant (p-
248 values ≥ 0.586; Fig. S24).

249 Based on the results above, the experiments with iNextEra libraries focused on baits
250 with 1/4 and 1/8th dilutions (Supplementary Notes #11.2; Table S15). The number of raw
251 reads obtained was similar for both bait dilutions and unenriched (Fig. S25). The
252 percentage of mapped reads was correlated with the bait input amount, averaging ~18%
253 for the enriched libraries and 0.11% for unenriched libraries (Fig. S26, p-value ≥ 0.337).
254 The depth and breadth of *C. parvum* genome coverage were, on average, 43.9X and
255 41.67% for libraries enriched with quarter dilutions of the baits, 11.8X, and 20.64% for
256 libraries enriched with an eighth dilution of the baits, vs. 0.03X and 1.55% for unenriched
257 libraries, however, values were not significantly different, even after normalization based
258 on retained reads (all p-values ≥ 0.091; Figs S27-28).
259
260 *Enriching Seven Human Fecal DNA samples with the CryptoCap_100k Bait Set*
261 For libraries from DNA obtained from seven clinical patients' fecal samples, roughly
262 a similar average number of reads were obtained for unenriched and enriched libraries
263 (Supplementary Notes #12.1, Table S16, Fig. S29). The percentage of reads mapping to
264 10-CrypGS increased from 0.02-0.15% (unenriched) to 0.37-18.17% (enriched), with a
265 fold change of 9.25-84.5X for the four DOCK samples and 354-479X for the three UK
266 samples (Table S16, Fig. S30, p-value = 0.02). Species assignment using 10-CrypGS
267 shows drastic differences between enriched and unenriched libraries (Fig. S31); all
268 clinical samples from the unenriched libraries show a high proportion of the few mapping
269 reads assigned to non-human infecting *C. muris* (Fig. S31 right), similar to what is
270 observed as background effect from samples with low concentration in the LOD
271 experiment. In contrast, DOCK and UKH101 enriched libraries yield reads assigned

272 mostly to *C. hominis* and *C. cuniculus*. Samples EC4 and UKP196 contain high
273 proportions of *C. parvum* (Fig. S31 left). No reads mapped to the *gp60* database, and very
274 few mapped to the *18S rRNA* database with assignment to species not typically observed
275 in human infections (Fig. S32).
276
277 *Enriching 100 Human Fecal DNA Samples with the CryptoCap_100k Bait Set*
278 iTru libraries
279 For libraries prepared from fecal DNA obtained from 100 patients with
280 cryptosporidiosis using the iTru protocol, 26 samples required more than one attempt and
281 an increased number of PCR cycles (from 14 to 18) at library preparation to produce a
282 library that passed QC. We obtained an average of 2.6 million paired reads per enriched
283 library for the 100 samples, and 4.1 million paired reads for 73 samples from unenriched
284 libraries (Supplementary Notes #12.2.1, Table S17; Fig. S33). On average, 35.6% of
285 enriched vs. 0.08% of unenriched retained reads mapped to the *C. parvum* genome,
286 respectively (p-value < 0.0001; Fig. 5A; Table S17; Fig. S34). The libraries were
287 enriched to *C. parvum* by an average factor of 2,095X (range 7.20 – 5,097X; Table S17).
288 The depth and breadth of genome coverage were significantly higher for enriched
289 samples with an average depth normalized per 100,000 reads of 1.12X and breadth of
290 2.25% for enriched vs. an average depth of 0.0004X and breadth of 0.0008% for the
291 unenriched libraries (p-values < 0.0001; Fig. 5B; Table S17; not normalized values Fig.
292 S35).
293 Next, we assessed species assignment from the relative number of hits from the iTru
294 libraries to each species in 10-CrypGS. The percentage of reads mapping was positively

295 but weakly correlated to the qPCR C_T values for unenriched libraries and negatively
296 correlated with enriched libraries (Figs. S36-37, Supplementary Results). A 5.6%
297 increase of paired reads mapping to 10-CrypGS vs. *C. parvum* genome was observed for
298 enriched libraries and 0.05% for unenriched (Table S17). Still, only 53 unenriched
299 samples mapped to the reference, with an average of 2,705 mapped reads per sample. The
300 species assignments for unenriched libraries generally resemble the complex background
301 used for the dilution experiments, including mostly hits to *C. muris* (cf. Fig. 4B and 5C).
302 Following a single round of enrichment, an average of 1.03 million reads hit the database
303 (Table S17), with most hits matching *C. parvum* (63 samples have >50% of hits to *C.*
304 *parvum*; Fig. 5D), but many samples have moderate proportions of reads assigned to
305 other species, including species that have limited evidence that they infect humans (Fig.
306 5D; see Supplementary Results). Thus, species assignments differed drastically between
307 enriched vs. unenriched samples (Fig. 5C vs. 5D), with a significant shift in species
308 assignments toward species with more evidence of human infection, except that a slightly
309 lower proportion of reads are assigned to *C. hominis* (see Supplementary Results).
310
311 <u>iNextEra Libraries</u>
312 For the 91 clinical samples for which iNextEra protocol was used, we obtained an
313 average of 3.7 million paired reads per unenriched library, 5.3 million paired reads for
314 single-enriched libraries, and 3.6 million for double-enriched libraries (Supplementary
315 Notes #12.2.2, Table S18; Fig. S38). On average of 0.03% of unenriched reads, 15.85%
316 of single-enriched reads, and 64.98% of double-enriched reads mapped to the *C. parvum*
317 genome (Supplementary Notes #12.2.2, all p-values < 0.0001, Fig. 6A, Table S18, Fig

318 S39), yielding an average enrichment factor of 1,112X from unenriched to single-
319 enriched, and an extra 9.24X from single- to double-enriched (Table S18). The depth and
320 breadth of genome coverage normalized per 100,000 reads were significantly higher for
321 double-enriched samples, followed by single-enriched samples (Fig. 6B). The depth of
322 genome coverage per 100,000 reads was for unenriched = 0.001X, single-enriched =
323 0.563X, and for double-enriched = 2.70X. The breadth of genome coverage was for
324 unenriched = 0.060%, for single-enriched = 1.79%, and for double-enriched libraries =
325 5.99% (Fig. 6B, Table S18, not normalized values Fig. S40). The percentage of reads
326 mapping to the 10-CrypGS was negatively correlated to the qPCR C_T values for all
327 library types. Double-enrichments allow retrieving a higher percentage of target reads at
328 higher Ct values than the other two library types (Figs. S41-S43, Supplementary Results).
329 Species assignments were similar to those from iTru libraries in unenriched and single-
330 enriched libraries (cf. Figs. 5C vs. 6C, and Figs. 5D vs. Fig. 6D). The species assignment
331 analysis shows a substantially increased proportion of hits to the most frequently
332 observed human-infecting species, *C. parvum,* and *C. hominis* in double-enriched
333 libraries compared to single- and unenriched libraries (Fig. 6E).
334
335 <u>iNextEra vs. iTru Libraries</u>
336 iNextEra and iTru libraries performed similarly (Table S19) but with noteworthy
337 differences. For single-enriched libraries, we obtained an average of 2.04X more read
338 pairs for iNextEra than iTru libraries (Supplementary Notes #12.2.3, Table S19; Fig.
339 S44). More reads were retained following quality filtering for iNextera vs. iTru, but this
340 result is correlated with sequencing instruments used (iNextEra used NovaSeq X vs. iTru

341 used HiSeq X). A higher percentage of reads mapped to the *C. parvum* reference in
342 samples from iTru vs. iNextEra libraries (35.6% vs. 15.9%, respectively; p-value <
343 0.0001; Fig. S45), which resulted in iTru libraries yielding an average enrichment factor
344 about twice as large as iNextEra libraries (Table S19). However, the normalized depth of
345 coverage is very similar for both protocols (0.5X, p-value = 0.7718; not normalized
346 values Fig. S46). The normalized breadth of coverage was significantly higher for the
347 iNextEra libraries (p-value < 0.0001; not normalized values Fig. S47, both normalized
348 Fig. S48).
349

350 **Discussion**

351 Due to the limitations of currently available methods employed to study
352 *Cryptosporidium* spp., there is an overall lack of understanding of *Cryptosporidium*
353 biological and genomic diversity and how it correlates with transmission, infection,
354 pathogenesis, and host response. Here, we developed and critically evaluated a capture
355 system to obtain *Cryptosporidium* DNA from samples with low input DNA levels,
356 varying target concentrations, and varying species composition to generate genomic data
357 from human-infecting *Cryptosporidium* species. Our bait capture assay increases on-
358 target reads by 2-3 orders of magnitude, enables the pooling of ~10 libraries per capture
359 reaction, and the dilution of the baits, thus significantly increasing the affordability of use
360 in future studies (see details below). Our study revealed several significant findings: (i)
361 CryptoCap_100k is highly specific and achieves a higher percentage of on-target reads
362 than unenriched libraries in clinical samples with both iTru (~437X increase) and iNext
363 (~500X in single-enriched and ~2,167X in double-enriched) library preparation methods,

364 (ii) in DNA samples from pure oocysts or manufacturer, CryptoCap_100k can cover 99%
365 of the genome without deep sequencing, and (iii) in real-world human fecal DNA
366 samples, CryptoCap_100k recovers sequences for an average of about half of the
367 genome, but with a relatively wide range, including samples with high Ct values, and (iv)
368 double enrichments increase the number of on-target reads even further and provide
369 substantially better species assignment.

370 This study was not the first to apply bait capture to whole genome sequencing of
371 human-infecting *Cryptosporidium* spp. (Khan and Alves-Ferreira et al. unpublished).
372 However, our study is the first to design baits based on new and more complete genome
373 sequences (e.g., *C. parvum* IOWA-ATCC[41]) and to include baits to account for variation
374 among species of human-infecting *Cryptosporidium* spp. by selecting for unique regions
375 of these genomes when compared to the *C. parvum* genome sequence. Thus, our new bait
376 set could yield sequences from variable regions within and among species of
377 *Cryptosporidium* while being highly specific, as shown by the low levels of cross-
378 mapping to other non-*Cryptosporidum* species. In addition, our study innovates using *in*
379 *silico* testing to evaluate baits on multiple species of *Cryptosporidium* and provide
380 specific details on the expected performance of the hybridization bait capture assay. In
381 clinical samples, our study finds that CryptoCap_100k provides a higher percentage of
382 target reads (Fig.5A and 6A) and higher depth and breadth of genome coverage (Fig. 5B
383 and 6B) across all samples compared to unenriched libraries. These findings are
384 comparable to our *in silico* results (Table S1, Fig. 1). Thus, the CryptoCap_100k bait set
385 can cover 99% or more of the *Cryptosporidium* spp. genome *in silico* and ~ 50% *in vitro*

386 in samples single-enriched with iNextEra with Ct values ranging from 18.01 to 31.08
387 (Table S19).
388 Studies on the population genetic structure of *Cryptosporidium* spp. are needed, but
389 obtaining pure and sufficient *Cryptosporidium* DNA for validation and testing is
390 challenging[42]. Purifying oocysts can be labor-intensive and time-consuming, require
391 specialized equipment, and may result in significant loss of oocysts and DNA[43]. As one
392 of our goals, we aimed to remove the need to purify oocysts to obtain usable genomic
393 data. To make recommendations about how to best use our CryptoCap_100k baits, we
394 define two distinct limits of detection (LOD): (i) the LOD for minimum target genomic
395 equivalents and (ii) the LOD for the minimum bait concentration used in the assay. We
396 diluted our pure *C. parvum* DNA in a complex background community to reflect accurate
397 environmental or clinical samples and determine the threshold at which we no longer
398 obtain quality data for our enrichment libraries. Departing from the fact that a single
399 haploid sporozoite contains a little over ~9 fg of DNA. Our LOD for *C. parvum* assay
400 using both the NEB and iNextEra enriched libraries found that we obtained a high
401 percentage of target reads, substantial breadth and depth of genome coverage, and
402 accurate species assignment with inputs as low as 0.01 ng (~1,000 sporozoites or ~250
403 oocysts equivalents) of *C. parvum* DNA (Figs. 4A-4B). Moreover, we were able to
404 improve all these parameters further in double-enriched iNextEra libraries that contained
405 0.001 ng (~100 sporozoite or ~25 oocyst equivalents) of *C. parvum* DNA (Fig. 4D). In
406 comparison, the unenriched libraries for both the NEB and iNextEra preparation methods
407 had extremely poor species assignment when input was below 0.1 ng (Fig. 4B & Fig.
408 4D). These findings establish that our CryptoCap_100k baits can obtain ample genomic

409 information at low levels, thus eliminating the need for purifying oocysts to gain useful
410 information. When assessing the LOD for the CryptoCap_100k baits, we found that there
411 was no significant difference in the percentage of reads that mapped, depth and breadth
412 of genome coverage observed between the different bait dilutions (full down to a
413 sixteenth, Figs. S24-S25), still we observed a trend where a higher percentage of target
414 reads and breadth were obtained at quarter dilutions and higher (Figs. S26-S27). Our
415 findings demonstrate that the baits can be diluted to save on cost.

416 Clinical human stool samples provided for testing are characterized by a small
417 volume containing a low overall abundance of *Cryptosporidium* oocysts in a large
418 background community of non-target organisms. This problem is heightened by the lack
419 of a readily available *in vitro* culture system[42,44,45]. We successfully tested our
420 CryptoCap_100k baits on an initial small set of seven samples. Once the approach
421 worked well (Fig. S31), we proceeded with the larger set of 100 clinical samples. We
422 observed significantly more reads mapping to *Cryptosporidium* genome sequences in
423 enriched vs. unenriched libraries with both the NEB and iNextEra library methods (Fig.
424 5A; Fig. 6A). Furthermore, the average breadth and depth of genome coverage for the
425 100 clinical samples were significantly higher for both enriched library preparation
426 methods compared to unenriched (Fig. 5B; Fig. 6B). This demonstrates an increased level
427 of on-target reads for both library preparation methods, leading to better species
428 assignment of the reads (Figs. 5C-D, Figs. 6C-E). Several studies recommend only
429 sequencing samples with Ct values < 16[43,46]. However, even with an average large Ct
430 value of 25.4, we still see, on average, 31.3% breadth and 28.1X depth in the clinical
431 samples using NEB libraries and 54.38% coverage and 32.11X depth in the clinical

432 samples using iNextEra libraries (Table S17-S18). Our linear regression model estimated
433 that the percentage of reads mapped against 10-CrypGS decreased by only 2.8% for each
434 increment of Ct units for NEB-enriched libraries (Fig S36). For iNextEra libraries,
435 unenriched libraries show 0% target reads at a Ct value of 35 (Fig. S41), and single-
436 enriched libraries show 0% target reads approx. at 40 Ct (Fig. S42), and double-enriched
437 libraries display a flatter trend, indicating that even high Ct-value samples tend to retain
438 high percentages (Fig. S43). Following this trend, double enrichments with iNextEra
439 library preparation always resulted in significant percentages of reads mapped to the *C.*
440 *parvum* genome and the depth of the genome covered (Fig. 6A-B), as well as more
441 accurate species assignment from the reads obtained (Figs. 4D & 6E). Species assignment
442 following enrichment coincides with the species detected in these same samples using *C.*
443 *hominis* and C. *parvum* species-specific duplex real-time PCR assay[47]. Double
444 enrichment could be especially useful when seeking accurate information from samples
445 believed to be a new strain or species or to increase the confidence in variations found in
446 the genome by increasing the depth of sequencing coverage. Additionally, the increase in
447 on-target reads and depth could aid researchers in parsing mixed infections.

448 Clinical *Cryptosporidium* infections can contain multiple strains, subtypes, or species
449 within one infection. Current methods for detecting mixed infections include amplicon,
450 PCR-RFLP, and MLST approaches[21,48,49]. However, these methods have inherent biases
451 due to preferential amplification and can be difficult to execute for mixed infections that
452 are not comprised of *C. parvum* and *C. hominis*[21,22,48]. We simulated *in vitro* a mixed
453 infection using two DNA samples of *C. parvum* and *C. meleagridis* pooled at planned
454 equal proportions. We found a slightly higher proportion of *C. parvum* than *C.*

455 *meleagridis* in unenriched and enriched reads. These differences may be due to pipetting
456 errors or discrepancies in the abundance of the target in the two DNA samples, thus
457 creating unequal competition. However, we demonstrated that in the case of a mixed
458 infection, two species of *Cryptosporidium* were captured and identified, though the subtle
459 increase in the proportion of *C. parvum* in the enriched pool could also suggest that
460 enrichment could be slightly biased (Fig. 3A). This could be expected as *C. parvum* was
461 the base species used for bait design, however, such biases were not conspicuous in the
462 mixed-infection computer simulations (Fig. 2) nor in clinical samples that showed *C.*
463 *hominis* assignments. Potential mixed infections were detected in EC4 (Fig. S31) and
464 UKH151 (Fig. 6E). Overall, our CryptoCap_100k baits can readily detect mixed species
465 infections and can obtain estimates without extreme skewing of the relative abundances,
466 except perhaps for *C. hominis* and *C. cuniculus*, in which their introgression history
467 makes it challenging when using our current 10-CrypGS database. Further updates of
468 databases for species assignment are necessary tools for the community to help overcome
469 this issue. Our bait set provides the ability to generate data that could be used not only for
470 insights into the genome sequences of lesser sequenced species or subtypes but also to
471 populate these databases and refine analyses.

472 Our assay can be implemented in various scenarios, and the cost savings will vary
473 depending on the abundance of *Cryptosporidium* DNA in the sample. To illustrate costs,
474 please see Supplemental Results and Table S20. In brief, our assay reduces the cost of
475 characterizing genome-wide *Cryptosporidium* DNA sequences in fecal samples by an
476 average of about 98% (Table S20, Supplementary Results).

477 In conclusion, our findings demonstrate the ability of the CryptoCap_100k bait set
478 to cover >80% of the genome with high depths *in vitro* in pure samples (Figs. S13-S14)
479 and achieve as high as 3,690X fold-change enrichment in clinical samples (Table S18),
480 and as a result high accuracy of species assignment. Studies using the CryptoCap_100k
481 bait set can improve our understanding of the population genetics of human-infecting
482 *Cryptosporidium* species and their association with epidemiology. Finally, generating
483 economical and efficient genome-level data allows us to grow our databases and refine
484 analyses and methods.
485

486 **Methods**

487 *CryptoCap_100k Bait Design & Characterization*

488 Based on sequence similarity and availability on CryptoDB.org, six species'
489 genomes of *Cryptosporidium* were chosen for bait design (*C. parvum*, *C. hominis*, *C.*
490 *cuniculus*, *C. tyzzeri*, *C. meleagridis*, and *C. viatorum*), along with 55 *18S rRNA*
491 sequences and eight *gp60* sequences (see Data Availability). All sequences were soft-
492 masked for simple and low-complexity repeats using RepeatMasker v. 4.1.1
493 (http://www.repeatmasker.org/). From all sequences, we aimed to design 100,000 120-
494 nucleotide baits. Only baits that were ≦50% soft-masked for repeats were retained.
495 Applying the design configuration to all six genome sequences resulted in 435,323 raw
496 baits. Then, VSEARCH v. 2.15.0[50] was used to cluster baits at various levels of sequence
497 overlap and identity (Table 1). This approach used tolerance for mismatches between the
498 bait and target sequence to reduce the number of baits while maintaining diversity.
499 Furthermore, VSEARCH uses a "greedy algorithm," which results in bias based on the

500 order of the input sequences. This was used to the advantage of the design by selecting
501 the genome order for each bait set to be clustered to give preference to specific genome
502 sequences based on completeness and importance. The input order was *C. parvum*, *C.*
503 *hominis*, *C. cuniculus*, *C. tyzzeri*, *C. meleagridis*, and *C. viatorum*. *C. parvum* IOWA-
504 ATCC was used as the alignment reference (listed first) for all clustering procedures: *C.*
505 *hominis*, *C. cuniculus*, and *C. tyzzeri* were collapsed at 50% overlap, and 90% identity
506 relative to *C. parvum*; *C. meleagridis* and *C. viatorum* were collapsed at 45% overlap and
507 85% identity relative to *C. parvum*. Due to the high number of baits obtained from *C.*
508 *viatorum*, a subset was randomly removed. Raw 403 baits were obtained for the *18S*
509 *rRNA* targets; we collapsed these at 50% overlap and 95% identity to arrive at the final
510 set.

511 After bait design, we characterized the CryptoCap_100k bait set by using twelve
512 genome sequences available for ten species of *Cryptosporidium* (please see Data
513 Availability). BWA v 0.7.17[51] was used to map the CryptoCap_100k bait set to each
514 genome sequence. Alignment files were generated and converted using Samtools
515 v.1.10[51,52]. Then, the bamCoverage function from DeepTools v. 3.3.1[53] was used to
516 produce a bigwig file, which generated a coverage track calculated as the number of
517 reads per bin, where bins are short consecutive counting windows of a defined size of 50.
518 Then, bigwig files were visualized in IGV-Web app version 1.10.8[54]. To obtain alignment
519 statistics, the depth option from Samtools v1.1.0 and the *-a* flag were used to get all sites
520 and create a depth file[52]. A custom awk script (Supplementary Notes #1) obtained the
521 mean depth and proportion of baits mapped to each reference. The breadth of coverage

522 on each genome was obtained by estimating the number of bases covered in the
523 alignment over the reference length when setting a minimum coverage depth of 4.
524 We also compared the CryptoCap_100k bait set to the one developed by Khan
525 and Alves-Ferreira et al. (unpublished), here named CryptoCap_73k, which, according to
526 our analyses (Supplementary Notes #2), contains 68,966 baits of 120 nt. We estimated
527 the same parameters above using the identical genome sequences for this bait set, and we
528 made further statistical comparisons using Welch's two-sample t-test.
529
530 *In Silico Simulations of Enrichment with CryptoCap_100k & Sequencing*
531 In a simple infection scenario, we used the twelve available genome sequences for
532 ten *Cryptosporidium* species to simulate the enrichment and sequencing process with the
533 CryptoCap_100k bait set (Supplementary Notes #3). Each genome was used as if it
534 belonged to a sample containing only such a genome. First, we obtained the coordinates
535 of the above-mapped baits on each of the twelve genome sequences (*CryptoCap_100k*
536 *Bait Design & Characterization*). To simulate the hybridization process, we extended
537 those mapping positions to several bases to the left and right, assuming that the bait was
538 annealing to the centroid of larger fragments being captured. We performed two types of
539 extensions to understand the effect of mean library fragment size on enrichment and data.
540 We thus simulated the capture of fragments with two different size distributions. We took
541 the first mapping positions for all baits and extended each region for large fragments, 200
542 bases to the left and 320 bases (120 bait +200) to the right, to simulate a library insert
543 size of 520 bp. Also, we extended the coordinate where the bait mapped 40 bases to the
544 left and 160 bases (120 bait + 40) to the right to simulate a library insert size of 200 bp.

545 Then, the faidx option from Samtools was used to extract the two types of extended
546 sequences from each genome to create a fasta file[52].
547 To simulate the sequencing process from these "captured" fragments, the software
548 ART 2016.06.05[55] was used to reproduce ~200,000 paired-end Illumina 150 bp fastq
549 reads from each fasta. Each set of fastq files was mapped to the respective genome
550 sequence using BWA v 0.7.17[51], and we estimated depth and breadth coverage from
551 these reads on each genome and produced bigwig files for visualization of the results.
552 Then, we assessed and validated the simulated data species assignment accuracy
553 (Supplementary Notes #4). To do this, we used three reference databases. First, we
554 generated one fasta file containing ten *Cryptosporidium* genome sequences that served as
555 a reference (10-CrypGS, from the twelve genome sequences reference, we removed *C.*
556 *hominis* UdeA01 and *C. parvum* IOWA II). Second, we used a database of *gp60*
557 sequences, the most widely used genetic marker for subtyping *Cryptosporidium*, which
558 contains the different subtypes identified in human cases[15]. Third, we used a curated
559 CryptoDB.org database of *18S rRNA*, the second most characterized loci in the group[56–
560 58]. We used BBMap 38.90[59] to map fastq reads to each of the three reference files, which
561 by default allows multi-mapping (secondary alignment) of baits to the genomes. We only
562 considered regions where both reads were mapped to count the number of hits. We
563 plotted the proportion of paired hits to specific *Cryptosporidium* species for each
564 simulated dataset.
565 To understand the specificity of the bait set, by the sensitivity of the baits to
566 hybridize to off-target DNA, common in environmental and clinical samples, we assessed
567 *in silico* the affinity of our baits to various genome sequences, including microbial,

568 human, and cow sequences (Supplementary Notes #5 and #6). To do this, we used a fasta
569 file with 1,008 whole genome sequences[40], which includes genomes of bacteria, archaea,
570 and eukaryote organisms, including the human genome. Separately, we also used the *Bos*
571 *taurus* genome sequence as a reference (GCA_002263795.3) to assess if our baits can
572 hybridize to DNA used as background for dilution *in vitro* experiments. For both
573 analyses, metagenomes, and *Bos taurus*, we used BWA 0.7.17[51] to map the baits. Then,
574 we estimated the number of hits as well as the quality of the match.

575 Finally, to explore *in silico,* the ability to discriminate different species within a
576 mixed-infection sample enriched with our bait set, we created four input files to simulate
577 different mixed infections scenarios (Supplementary Notes #7). We did this by
578 concatenating the fastq files already generated for each species above and using only
579 those designed large insert fragments. The four input types were: i) a fastq file containing
580 reads from all twelve genomes (*C. andersoni, C. baileyi, C. cuniculus, C. hominis* x 2*, C.*
581 *meleagridis, C. muris, C. parvum* x 2*, C. tyzzeri, C. ubiquitum,* and *C. viatorum*); ii) a file
582 containing the sequences of the ten representative species (*C. andersoni, C. baileyi, C.*
583 *cuniculus, C. hominis, C. meleagridis, C. muris, C. parvum, C. tyzzeri, C. ubiquitum,* and
584 *C. viatorum*); iii) a file containing the cluster of five species that share high percent
585 identity (*C. parvum, C. hominis, C. tyzzeri, C. meleagridis,* and *C. ubiquitum*); and 4) a
586 file containing the three most common species found in human infections (*C. hominis, C.*
587 *cuniculus,* and *C. parvum*). Then, we mapped these mixed reads to 10-CrypGS, the *gp60*
588 database, and the *18S rRNA* database following the same parameters as above.

589

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

591 Pure oocyst DNA was obtained for three *C. parvum* samples from Dr. Michael
592 Grigg (National Institute of Health, US) and a duplicate of *C. meleagridis* DNA isolate
593 TU1867 from BEI Resources (BEI Resources NR2521 Manassa, VA). Metagenomic
594 libraries were prepared with the New England BioLabs Ultra FS II library kit (please read
595 *Metagenomic Libraries for In Vitro Experiments*), and an aliquot of these were enriched
596 with CryptoCap_100k bait sets (please read *Hybridization Capture Enrichments*).
597 Enriched and unenriched libraries were sequenced in an Illumina MiSeq using a PE250
598 kit.
599
600 *In Vitro Simulation of Mixed Infections*
601 To test the ability to detect mixed infection samples, after NEB library
602 preparation of DNA samples (please read *In Vitro Enrichment & Validation with*
603 *CryptoCap_100k Baits with Pure Oocysts' DNA*) and before hybridization bait capture,
604 two library products were pooled to simulate a mixed infection. To accomplish this, one
605 indexed library product from a pure sample from *C. parvum* and one with different
606 indexes from *C. meleagridis* were pooled at equimolar ratios. After pooling, we divided
607 the products into two aliquots, enriched one with CryptoCap_100k baits, and left the
608 second unenriched. Then, cleaning and sequencing proceeded, as explained below. We
609 predicted that the number of reads obtained based on the demultiplexing based on
610 primers attached before enrichment should be roughly 50% for each species for each
611 sample. These pooled libraries were sequenced in an Illumina MiSeq using a PE250 kit.
612
613 *In Vitro Test of Limits of Detection with Different Levels of C. parvum Dilution*

614 To determine the sensitivity or limits of detection (LOD) of *Cryptosporidium*
615 when using CryptoCap_100k bait set to mimic samples with a low concentration of target
616 DNA, we performed two library preps, one experiment with NEB libraries and iTru
617 primers[60] and one experiment with iNextEra libraries and iNext primers[60].
618 For the experiment with NEB libraries, we diluted pure *C. parvum* genomic DNA
619 in a complex background community, such as a community of DNA molecules from calf
620 thymus (Sigma-Aldrich, St. Louis, MO). For this experiment, a 45-ng aliquot of calf
621 thymus DNA was pipetted into five tubes. In the first tube, 5 ng of *C. parvum* genomic
622 DNA was added. Following this, three 1:10 serial dilutions were performed using the first
623 tube as a starter. The fifth tube corresponded to a negative control, containing just the calf
624 thymus. Then, metagenomic libraries with the NEB protocol and enrichments were
625 prepared for these libraries (please read *Metagenomic Libraries for In Vitro Experiments*
626 and *Hybridization Capture Enrichments*). Number of reads obtained, percentage of reads
627 mapped to the corresponding genome, and depth and breadth of genome coverage were
628 estimated.
629 For the experiment with iNextEra libraries, we diluted genomic DNA from pure
630 preparations of *C. parvum* into the complex background community. For this experiment,
631 45 ng aliquots of calf thymus DNA were pipetted into six tubes. In the first tube, 10 ng of
632 *C. parvum* genomic DNA was added. Four additional 1:10 serial dilutions were
633 performed using the first tube as a starter. The sixth tube corresponded to the negative
634 control, containing just the calf thymus. We grouped the six tubes in five different
635 treatments: one unenriched, one single-enriched with one-quarter of the recommended
636 baits, one single-enriched with one-eight of the recommended baits, one double-enriched

637 with one-quarter of the recommended baits, and one double-enriched with one-eight of
638 the recommended baits. Enrichments were prepared according to *Metagenomic Libraries*
639 *for In Vitro Experiments* and *Hybridization Capture Enrichments*. The number of reads
640 obtained, percentage of reads mapped to the corresponding genome, depth and breadth of
641 genome coverage, and species assignment based on the mapping of reads against 10-
642 CrypGS were performed to allow comparisons between dilutions.
643
644 *In Vitro Test of Limits of Detection with Different Levels of Bait Dilution*
645 To determine the sensitivity or limits of detection (LOD) of *Cryptosporidium*
646 when enriching libraries with the CryptoCap_100k bait set, we modulated the quantity of
647 baits added to the capture reaction, as above, in one experiment with NEB libraries and
648 one experiment with iNextEra libraries.
649 For both NEB libraries and iNextEra libraries, the bait dilution sensitivity assay
650 performed considered pools of the dual-indexed metagenomic libraries from the dilutions
651 experiment for *Cryptosporidium* plus the negative control (please see *In Vitro Test of*
652 *Limits of Detection with Different Levels of C. parvum Dilution*). For NEB libraries, we
653 tested five different bait quantities: full volume (5.5 µL, per the manufacturer's
654 specifications), half, quarter, eighth, and sixteenth. For iNextEra libraries, we tested
655 quarter and eighth bait dilutions. In these assays, the enrichment procedure was followed
656 according to the description in *Hybridization Capture Enrichments*, where the partial bait
657 quantities were achieved by serially diluting baits with water to a half, quarter, eighth,
658 and sixteenth of the concentration of the stock baits. The number of reads obtained, the

659 percentage of reads mapped to the corresponding genome, and the depth and breadth of
660 genome coverage were compared.
661
662 *In Vitro Enrichment & Validation with CryptoCap_100k Baits with Clinical Samples*
663 Two clinical datasets were used for real-world validation steps. First, Dr. Michael
664 Grigg (National Institute of Health, US) and Prof. Rachel Chalmers's laboratory (Public
665 Health Wales Microbiology and Health Protection, United Kingdom) provided DNA
666 extracted from seven de-identified patient fecal samples. These seven samples were
667 obtained on a NIAID IRB-approved protocol NCT00006150. The seven clinical samples
668 were extracted using the Qiagen PowerSoil Kit (Qiagen, Hilden, German). For this set of
669 samples, NEB library protocol was used, and aliquots were enriched with
670 CryptoCap_100k bait sets (please read *Hybridization Capture Enrichments*). Enriched
671 and unenriched libraries were sequenced in an Illumina HiSeq using a PE150 kit (300
672 cycles).
673 Second, 100 DNA extractions (Qiagen QIAamp Fast DNA Stool kit) from clinical
674 samples from Prof. Rachel Chalmers's laboratory (Public Health Wales—PHW—
675 Microbiology and Health Protection, United Kingdom) were used. This work was carried
676 out in accordance with a material transfer agreement between UGA and PHW for the
677 analysis of de-identified *Cryptosporidium* DNA, which did not require ethical approval.
678 A real-time qPCR assay was used to quantify the detection of *Cryptosporidium* species
679 from these clinical samples, using primers for the *18S rRNA* gene and probes slightly
680 modified from[61]. Here, we used as forward primer 5′ -
681 GGGTTGTATTTATTAGATAAAGAACCA- 3′, reverse 5′-

682 AGGCCAATACCCTACCGTCT-3′, and an internal probe 5′-
683 TGACATATCATTCAAGTTTCTGAC-3′. Amplifications were carried on a final
684 volume of 20 µL, which included 5 µL of 4x TaqPath™ 1-step RT-qPCR Master Mix
685 (Thermo Fisher Scientific, MD, USA), 1 µL of 10 µM forward primer, 1 µL of 10 µM
686 reverse primer, 2 µL of 2.5 µM fluorescent probe, 5 µL of 1:5 diluted DNA, and 6 µL of
687 molecular grade water. Two genomic library protocols were prepared for each (NEB and
688 iNextEra; please read *Metagenomic Libraries for In Vitro Experiments*). For both sets
689 and types of libraries, aliquots of these libraries were enriched with CryptoCap_100k bait
690 set; specifically, for iNextEra libraries, we also performed additional double enrichments
691 (please read *Hybridization Capture Enrichments*). Enriched and unenriched libraries were
692 sequenced in an Illumina NovaSeq using a PE150 kit.
693
694 *Metagenomic Libraries for In Vitro Experiments*
695 Metagenomic libraries were prepared using two methods. First, New England
696 BioLabs (NEB) Ultra FS II library kit (New England BioLabs, Ipswich, MA) was used.
697 Enzymatic fragmentation of DNA extracts was performed according to the manufacturer,
698 and it lasted between 5 to 11 minutes, based on the integrity of DNA visualized in a gel.
699 Ligation products were cleaned in 1X SpeedBeads[62]. A 12- or 14-cycle PCR was used to
700 add unique dual-indexes to each sample with iTru primers[60]. Samples were quantified
701 with a Qubit 2.0 Fluorometer DNA high-sensitivity assay kit (Thermofisher, Waltham,
702 MA) and cleaned with 1.25x SpeedBeads.
703 We implemented iNextEra library preparation protocols to decrease the time, cost,
704 and amount of input DNA needed to produce libraries that would pass QC. Nine samples

705 were depleted when producing iTru libraries and had no DNA remaining for iNextEra
706 libraries. The switch from iTru to iNextEra also coincided with a switch in Illumina
707 sequencing instruments, with the iNextEra libraries being run on the newly released
708 Illumina NovaSeq X. For those samples prepared following iNextEra Library Prep, in
709 brief, sample DNA was normalized to 10 ng/µL, for which 2.5 µL (25 ng) was used for a
710 tagmentation reaction that included 3 µL of 2X TMP buffer and 0.5 µL of Transposome
711 (BLT; CAT # 20015880, Illumina Inc.), incubated in a thermocycler at 53 °C for 30
712 minutes with the lid set at 80 °C. For PCR of the tagmentation products, a reaction
713 containing 9.5 µL of Q5 Reaction Buffer, 1.25 µL of dNTPs 10 mM, 0.5 µL of Q5 High
714 Fidelity Polymerase (New England Biolabs M0491L), 40.25 µL of nuclease-free water,
715 2.5 µL of iNext5 indexed primer at 5 µM, and 2.5 µL of iNext7 indexed primer at 5
716 µM[60], and 6 µL of tagmented product. The cycling conditions for PCR were: 3 minutes at
717 72 °C, 3 minutes at 98 °C, followed by ten cycles of 98 °C for 45 seconds, 62 °C for 30
718 seconds, and 72 °C for 2 minutes, with a final extension at 72 °C for 1 minute. Products
719 were validated in agarose gels and pooled with 5-7 other libraries at equimolar
720 concentrations. Pools were then cleaned with 1-1.25X SpeedBeads (Sigma Aldrich, St.
721 Louis, MO) and quantified with a Qubit 2.0 Fluorometer DNA (Thermofisher, Waltham,
722 MA). Cleaned pools were then aliquoted to have an aliquot of an unenriched pool and an
723 aliquot for enrichment with the CryptoCap_100k bait set.
724
725 *Hybridization Capture Enrichments with the CryptoCap_100k Bait Set*
726 For different experiments, the input concentration per pool varied (i.e., for clinical
727 samples using iTru/NEB, the input was between 25-80 ng/µL; for clinical samples using

728 iNext/iNextEra, it was 103 ng/µL; and for the LOD experiment it was between 26–59
729 ng/µL). The enrichment process in each pool followed manufacturers' protocols (mybaits
730 v5, Daciel Arbor Biosciences, Ann Arbor, MI), using the standard protocol, with wash
731 and hybridization temperatures of 63 °C and with modifications of bait quantities as
732 described above. After enrichment, A P5/P7 PCR of 16-22 cycles was performed,
733 followed by a 1-1.25X SpeedBeads clean-up (Sigma Aldrich, St. Louis, MO).
734 For the 100-clinical sample set and the LOD experiments, a second enrichment
735 (double enrichment) of the iNextEra single-captured pools was performed. For these,
736 aliquots of the cleaned single-captured pools were used independently as input for a
737 second capture that followed all manufacturers' steps from above, except that the volumes
738 and the amount of baits were brought to one-quarter of the recommended. For post-
739 double-enrichment P5/P7 PCR, 12-16 cycles were performed, followed by 1X
740 SpeedBeads clean-up. All PCRs were performed using the KAPA HiFi Kit (Roche,
741 Basel, Switzerland).
742
743 *In Vitro Bioinformatic Processing and Analyses*
744 Bioinformatic and data analysis code, including custom scripts, are available in
745 Supplementary Notes. All sequencing data were processed using Trimmomatic v. 0.39[63]
746 to remove Illumina TruSeq adaptors, low-quality bases and short reads. Trimmed
747 retained reads were then mapped to specific genomes or databases using BBMap 38.9[59].
748 Samtools[52] was used to sort bam files. Custom scripts were used to determine mapping
749 statistics such as genome depth and breadth of coverage and count the number of hits
750 against specific genomes.

751 Cleaned sequence reads from pure oocysts library samples and mixed-infection
752 samples were mapped to the corresponding target genome (i.e., pure *C. parvum* samples
753 to *C. parvum* IOWA-ATCC, pure *C. meleagridis* samples to chromosome-level
754 assembly[64], to 10-CrypGS, to the *gp60* database, and to the *18S rRNA* database (used in
755 simulations). The sequencing reads from the LOD dilution experiments (i.e., input DNA
756 and bait dilutions) were mapped to *C. parvum* and 10-CrypGS. The cleaned sequence
757 reads from clinical samples were mapped to *C. parvum* IOWA-ATCC and to 10-CrypGS.
758 Data analyses were performed in R Studio 2023.06.0. Least-square comparisons
759 of the means between experimental groups (e.g., unenriched and enriched, single-capture
760 vs. double-capture, etc.) were performed to calculate the p-value using a confidence
761 interval of 0.95 with the package: *lsmeans*[65], which provides a t-test output when
762 comparing two variables, or performs a Tukey method adjustment when comparing a
763 family of three or more estimates. Other packages used were *tidyverse*, *ggplot*, and
764 *plotly*[66]. Additionally, qPCR Ct values obtained were then used in the *parsnip 1.1.1* R
765 package[67] to define a regression model to predict the percentage of reads mapped against
766 10-CrypGS using a linear function.
767

768 **Acknowledgments**

769 T.G. and J.C.K. disclose support from the National Institutes of Health under
770 award number NIH R01AI148667. M.G. discloses support from the Intramural Research
771 Program of the National Institute of Allergy and Infectious Diseases (NIAID) at the
772 National Institutes of Health. We also thank Prisha Sharma for sample processing,

Christopher Fitzgerald for sample tracking and processing, Fiifi Agyabeng-Dadzie for sample processing, and Mustafa Nural for sample database help.

Author Contributions

N.J.B-V designed experiments, conducted all simulations and statistical analyses, guided bioinformatic analyses, wrote the manuscript, and contributed to obtaining funding. A.H.S. designed experiments, conducted experiments, conducted bioinformatic analyses, and wrote the manuscript. M.S.B. designed and conducted experiments, wrote the manuscript, and contributed to obtaining funding. A.K. contributed to experiments and analyses, edited the manuscript, and contributed to obtaining funding. R.P.B. contributed to experiments and analyses, edited the manuscript, and contributed to obtaining funding. K.N.P. conducted experiments and contributed to the manuscript. M.I.U.B. conducted experiments. B.B. designed and oversaw the production and QC of the baits and contributed to the manuscript. G.R. contributed samples and analyses to the manuscript. R.M.C. contributed samples and analyses to the manuscript. A-F.E. contributed samples and to protocol development. M.E.G. contributed experimental designs, samples, and analyses to the manuscript, and to obtaining funding. J.C.K. designed experiments and analyses, provided samples, contributed to the manuscript, and obtained funding. T.C.G. designed experiments, suggested simulations, contributed to analyses, co-wrote the manuscript, and obtained funding. All authors read and approved the final version of the manuscript.

Competing Interests

796 B.B. and M.B. are employed by Diacel Arbor Biosciences, which manufactures and sells
797 the baits described herein.
798
799 **Data Availability**
800 The six genome sequences that supported the CryptoCap_100k bait design were
801 from GenBank: GCA_015245375.1, GCA_004337795.1, GCA_001483515.1,
802 GCA_007210665.1, GCA_004337835.1, and GCA_001593445.1.
803 The *18S rRNA* sequence data that supported the CryptoCap_100k bait design
804 were from GenBank with the EU250845, LC089976, DQ288166, EF641022, GU951714,
805 AF108864, LC012016, GQ121020, HQ397716, KR296812, EF641014, FJ435960,
806 JX416368, KP099082, AY120903, AY120902, JX644908, AY324641, AY324638,
807 AF159113.1, JQ413438, AF108861, GQ121021, KP730304, AY737602, AY731235,
808 JX294358, AY268584, KR819168, AY120904.1, KU608308, AF112576, EU553551,
809 JF710247, KP004205, JQ002555, KT235702, AB909498.1, AF316630, AY504515,
810 AY504516, AF262325, GQ227479, EU827297, AF093500, AF093497, AY954886,
811 HM116385, DQ650344, KY490554, KC305650, EF547155, FJ769050, HM243547, and
812 AY524773 accession codes.
813 The *gp60* sequence data that supported the CryptoCap_100k bait design were
814 from GenBank with the FJ490060, FJ490087, FJ490058, JX412915, JX412926,
815 JX412925, KC204983, and KC204984 accession codes.
816 The twelve *Cryptosporidium* genome sequences used for CryptoCap_100k bait
817 set characterization and *in silico* simulations were from GenBank with the
818 GCF_00000165345.1, GCA_004337835.1, GCA_004337795.1, GCA_007210665.1,

819 GCF_001865345.1, GCA_001865355.1, GCA_001593455.1, GCF_000006515.1,
820 GCA_000165345.1, GCA_001483515.1, and GCA_002223825.1 accession codes *C.*
821 *hominis* UdeA01 and *C. parvum* IOWA-ATCC were removed for those analyses with
822 only ten genome sequences reference database (10-CrypGS).
823 Generated fastq files from all *in silico* and *in vitro* experiments for mock and
824 clinical samples are available in GenBank BioProject number PRJNA1061798.
825
826 **Computer Code**
827 The code used for all sections of this paper is available as Supplementary Notes.
828 Please note that the code includes segments used to create the main figures, and thus,
829 these figures, along with supplemental ones, are also presented in the supplemental with
830 the corresponding numbers used in the manuscript.

References

1. Bouzid, M., Hunter, P. R., Chalmers, R. M. & Tyler, K. M. *Cryptosporidium* Pathogenicity and Virulence. *Clin Microbiol Rev* **26**, 115–134 (2013).
2. Chalmers, R. M. & Katzer, F. Looking for *Cryptosporidium*: the application of advances in detection and diagnosis. *Trends Parasitol* **29**, 237–251 (2013).
3. 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).
4. Feltus, D. C. *et al.* Evidence supporting zoonotic transmission of *Cryptosporidium* spp. in Wisconsin. *J Clin Microbiol* **44**, 4303–4308 (2006).
5. Nader, J. L. *et al.* Evolutionary genomics of anthroponosis in *Cryptosporidium*. *Nat Microbiol* **4**, 826–836 (2019).
6. 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).
7. Sponseller, J. K., Griffiths, J. K. & Tzipori, S. The Evolution of Respiratory Cryptosporidiosis: Evidence for Transmission by Inhalation. *Clin Microbiol Rev* **27**, 575–586 (2014).
8. Gilbert, I. H. *et al.* Safe and effective treatments are needed for cryptosporidiosis, a truly neglected tropical disease. *BMJ Global Health* **8**, e012540 (2023).
9. 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).

10. GBD Results. *Institute for Health Metrics and Evaluation* https://vizhub.healthdata.org/gbd-results.

11. Cryptosporidiosis NNDSS Summary Report for 2019 | Water-related Topics | Healthy Water | CDC. https://www.cdc.gov/healthywater/surveillance/cryptosporidium/cryptosporidium-2019.html (2021).

12. 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).

13. 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).

14. Hunter, P. R. & Nichols, G. Epidemiology and clinical features of *Cryptosporidium* infection in immunocompromised patients. *Clin Microbiol Rev* **15**, 145–154 (2002).

15. Khan, A., Shaik, J. S. & Grigg, M. E. Genomics and molecular epidemiology of *Cryptosporidium* species. *Acta Trop* **184**, 1–14 (2018).

16. 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).

17. Feng, Y., Ryan, U. M. & Xiao, L. Genetic Diversity and Population Structure of *Cryptosporidium*. *Trends Parasitol* **34**, 997–1011 (2018).

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

19. 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).

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

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

22. 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).

23. 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).

24. 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).

25. 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).

26. 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).

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

28. 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).

29. 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).

30. 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).

31. 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).

32. 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).

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

34. 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).

35. 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).

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

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

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

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

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

41. Baptista, R. de P., Xiao, R., Li, Y., Glenn, T. C. & Kissinger, J. C. New T2T assembly of *Cryptosporidium parvum* IOWA annotated with reference genome gene identifiers. *bioRxiv* 2023.06.13.544219 (2023) doi:10.1101/2023.06.13.544219.

42. 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).

43. 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).

44. 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).

45. 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).

46. 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).

47. 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.

48. 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).

49. 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).

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

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

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

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

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

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

56. 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).

57. 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).

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

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

60. 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).

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

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

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

64. 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.

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

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

67. Kuhn, Max; Vaughan, Davis; Hvitfeldt, Emil. yardstick. tidymodels (2023).

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_73k 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_73k 68,966 baits (Khan and Alves-Ferreira et al., unpublished)		
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 of mapping *in silico* simulated enrichments with CryptoCap_100k baits of 150 bp fastq paired-reads for each of the twelve genomes assuming two input fragment size classes (200 bp and 520 bp). These mappings were only against each of the corresponding genomes.

Species Information		Fragment Size	# Simulated paired-reads	Mean read depth	Breadth of coverage
Species included in design	*C. parvum* IOWA II	~200 bp	255,486	8.42	99.1%
		~520 bp	255,332	8.42	99.6%
	C. parvum IOWA ATCC	~200 bp	256,947	8.45	99.4%
		~520 bp	256,947	8.45	99.9%
	C. hominis UdeA01	~200 bp	256,179	8.50	99.4%
		~520 bp	255,634	8.48	99.8%
	C. hominis 30976	~200 bp	257,208	8.52	99.6%
		~520 bp	257,102	8.51	99.9%
	C. cuniculus UKCU2	~200 bp	174,845	5.71	74.7%
		~520 bp	185,561	6.07	74.5%
	C. tyzzeri UGA55	~200 bp	249,737	8.31	98.4%
		~520 bp	248,498	8.27	98.9%
	C. meleagridis	~200 bp	241,764	8.07	98.5%
		~520 bp	241,770	8.08	99.9%
	C. viatorum ABER_CVIA_1.0	~200 bp	243,998	7.92	83.7%
		~520 bp	243,985	7.92	97.4%
Species not included in design	*C. andersoni* 30847	~200 bp	209,400	7.37	1.4%
		~520 bp	214,247	7.53	3.3%
	C. baileyi TAMU-09Q1	~200 bp	199,404	7.10	7.0%
		~520 bp	200,042	7.12	15.3%
	C. muris RN66	~200 bp	198,254	6.39	1.8%
		~520 bp	197,720	6.49	3.7%
	C. ubiquitum 39726	~200 bp	198,484	6.65	42.6%
		~520 bp	198,485	6.65	69.0%

Table 4. 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 S17 and S18.

	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 genomes	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 not-enriched counterparts, which is not displayed in Tables S17 and S18.

Figures

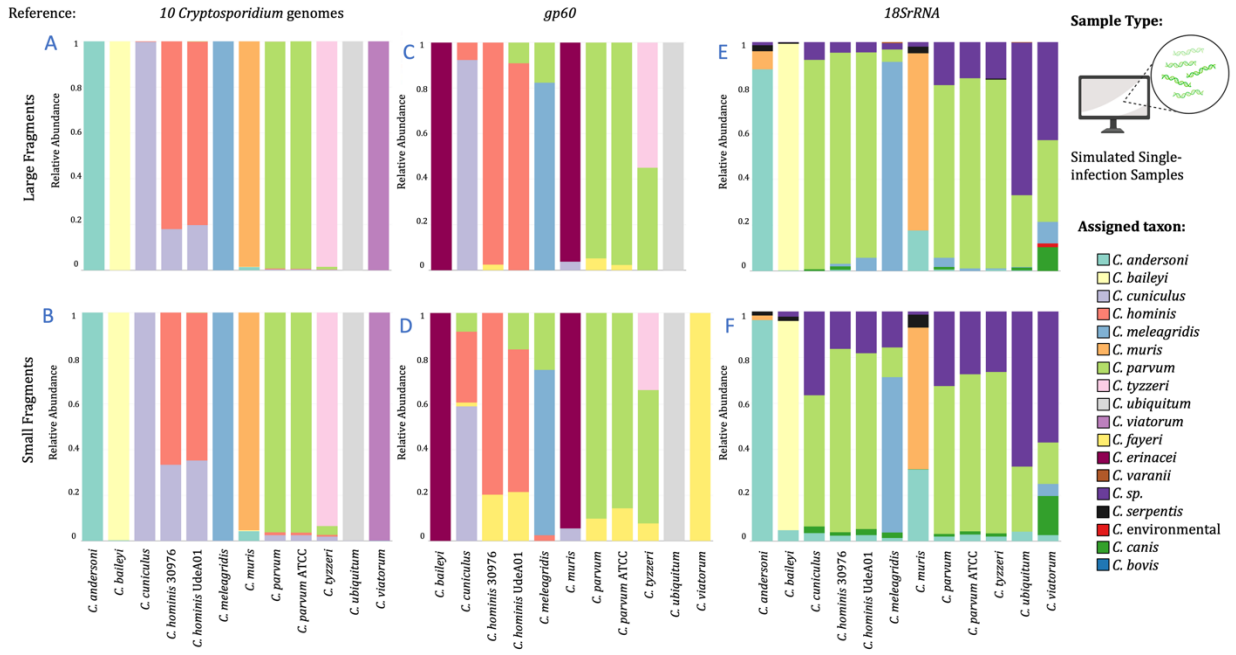


Figure 1. Barplots of relative abundance of species assignment from mapping *in silico* simulated enriched and sequenced reads for each of the twelve *Cryptosporidium* genomes to three reference databases. Simulations were modeled separately for two fragment-insert size classes, one of large-520 bp fragments (above) and one of small-200 bp fragments (below). Three reference databases were used to map simulated reads, one containing ten *Cryptosporidium* genome sequences (10-CrypGS), a *gp60* database (Khan et al., 2018), and an *18S rRNA* database (Xiao et al., 2010).

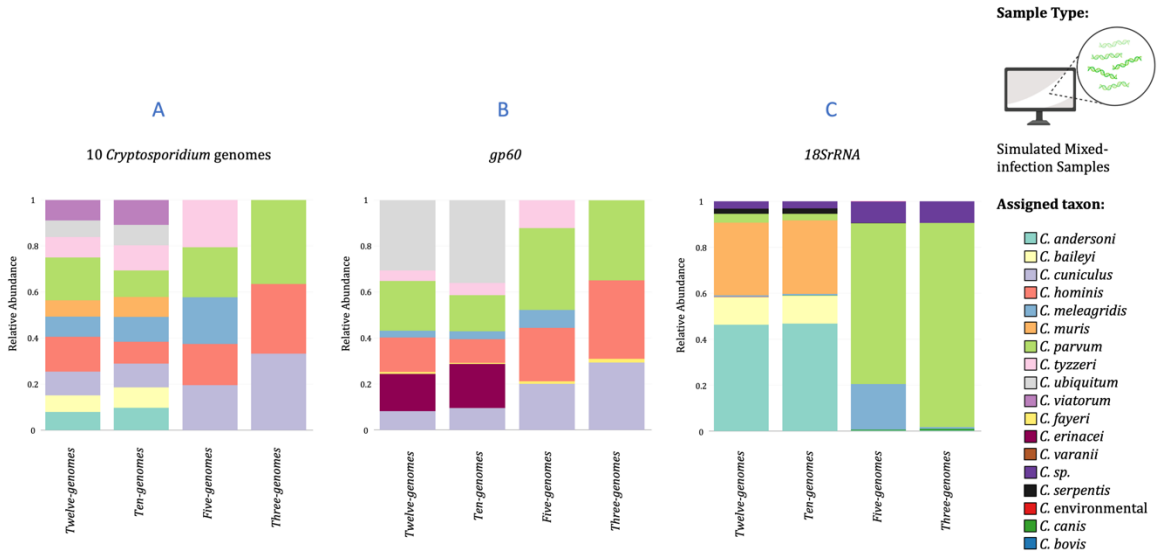


Figure 2. Barplots of relative abundance of species assignment from mapping four classes of *in silico* simulated enriched and sequenced mixed-infection samples. Three reference databases were used to map mixed infection simulated reads, one containing ten *Cryptosporidium* genome sequences (10-CrypGS), a *gp60* database (Khan et al., 2018), and an *18S rRNA* database (Xiao et al., 2010).

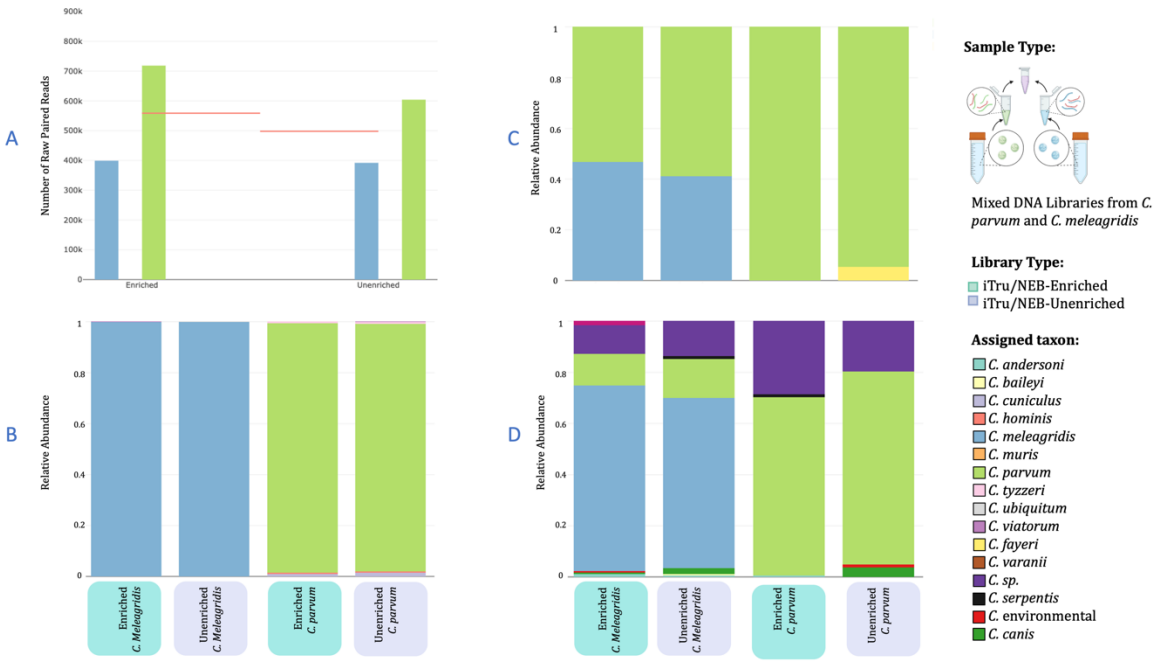


Figure 3. Species assignment plots from mapping one mock mixed-infection sample for which we sequenced unenriched and enriched libraries. (A) The number of raw paired reads obtained for each species in the mix, *C. parvum* in green and *C. meleagridis* in blue for both library types. The pink lines indicate 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.

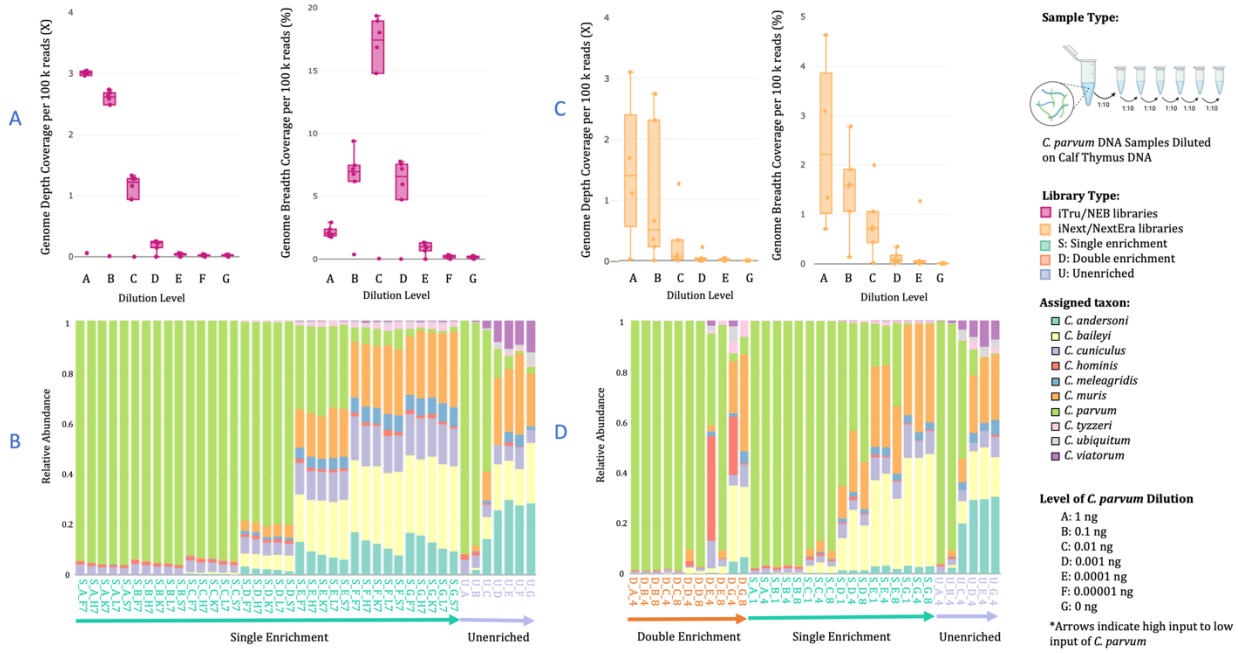


Figure 4. Summary statistics for sequence data from serially diluted *C. parvum* samples to assess detection levels given different input amounts. (A) *C. parvum* genome depth and breadth for diluted DNA samples prepared with iTru/NEB libraries normalized by 100,000 reads, (B) species assignment of diluted *C. parvum* samples prepared with iTru/NEB 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 for diluted DNA samples prepared with iNext/NextEra libraries normalized by 100,000 reads, (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).

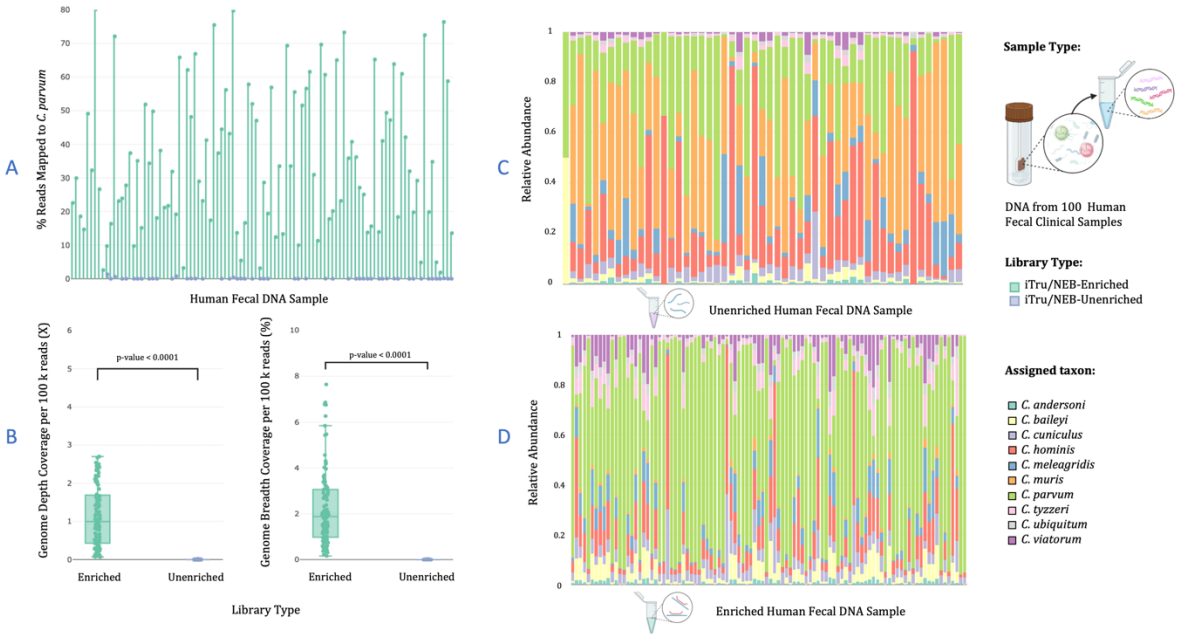


Figure 5. Summary statistics for sequence data from 100 clinical samples set prepared with the NEB protocol and iTru primers, comparing libraries enriched with CryptoCap_100k bait set versus unenriched libraries for validation. (A) Scatter plot of percentage of reads mapped from one round of enrichment 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 mapped regions for each sample in enriched libraries compared to unenriched normalized by 100,000 reads, and (C) species assignment to each sample prepared with NEB protocol and enriched when mapped to a reference database containing ten *Cryptosporidium* genome sequences (10-CrypGS). (D) species assignment to each sample prepared with NEB protocol and left unenriched when mapped to a reference database containing ten *Cryptosporidium* genome sequences (10-CrypGS).

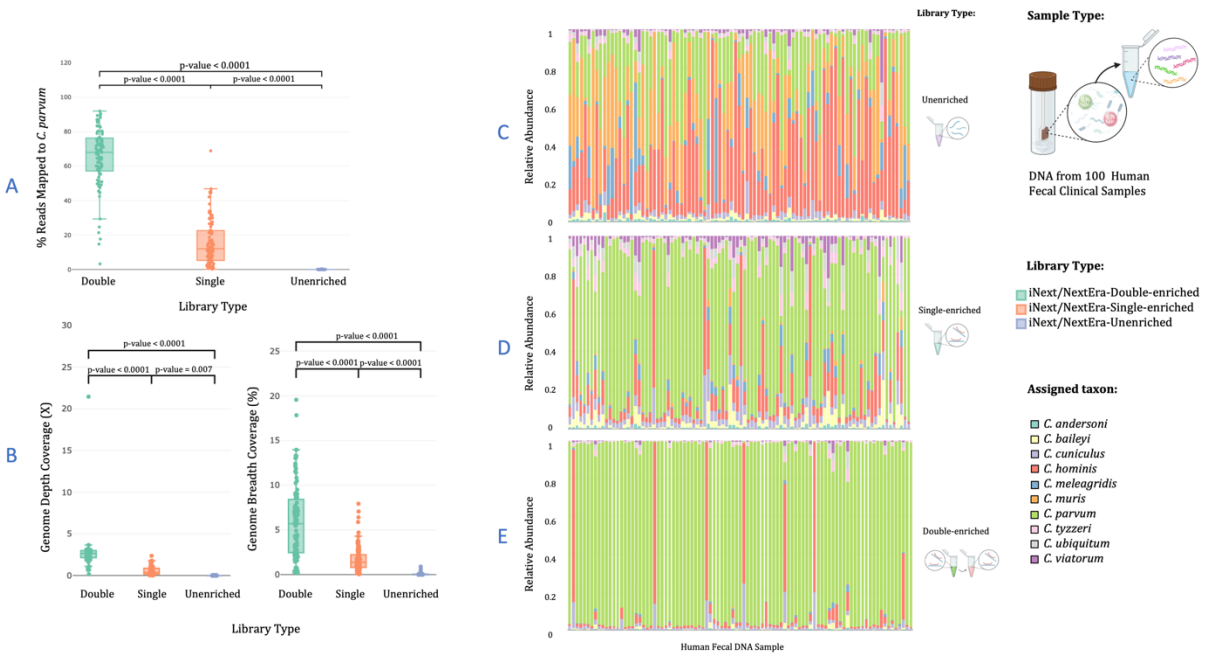
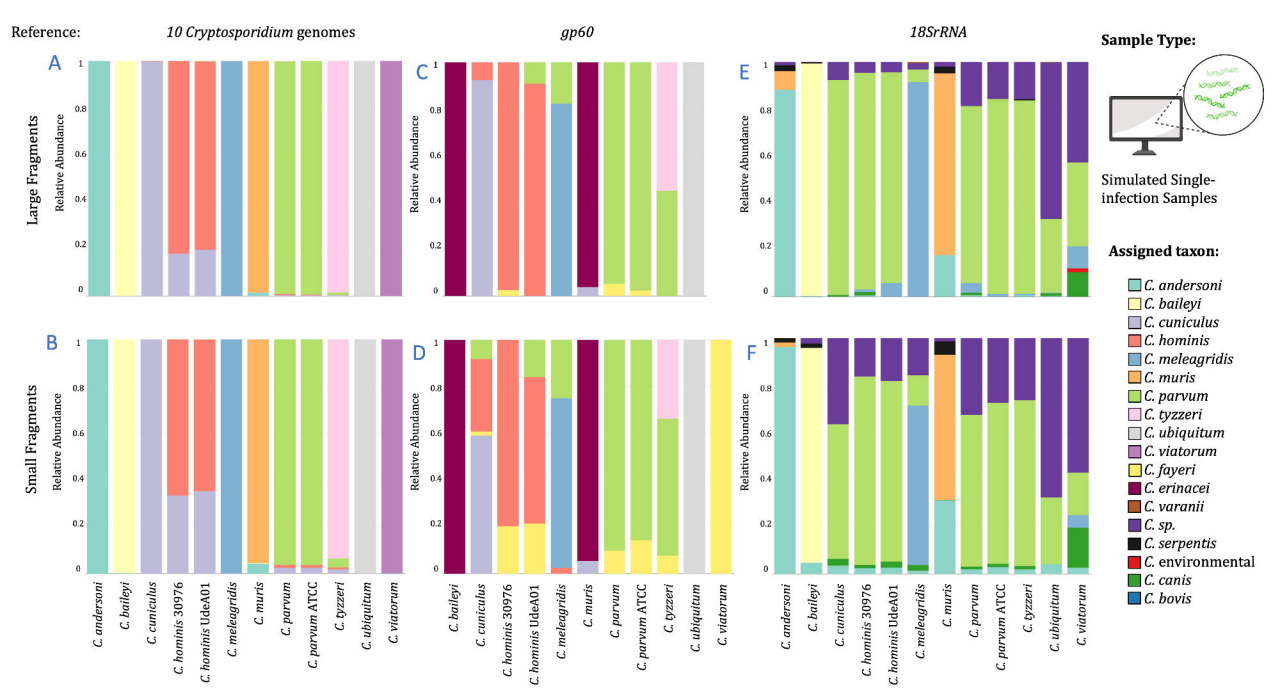
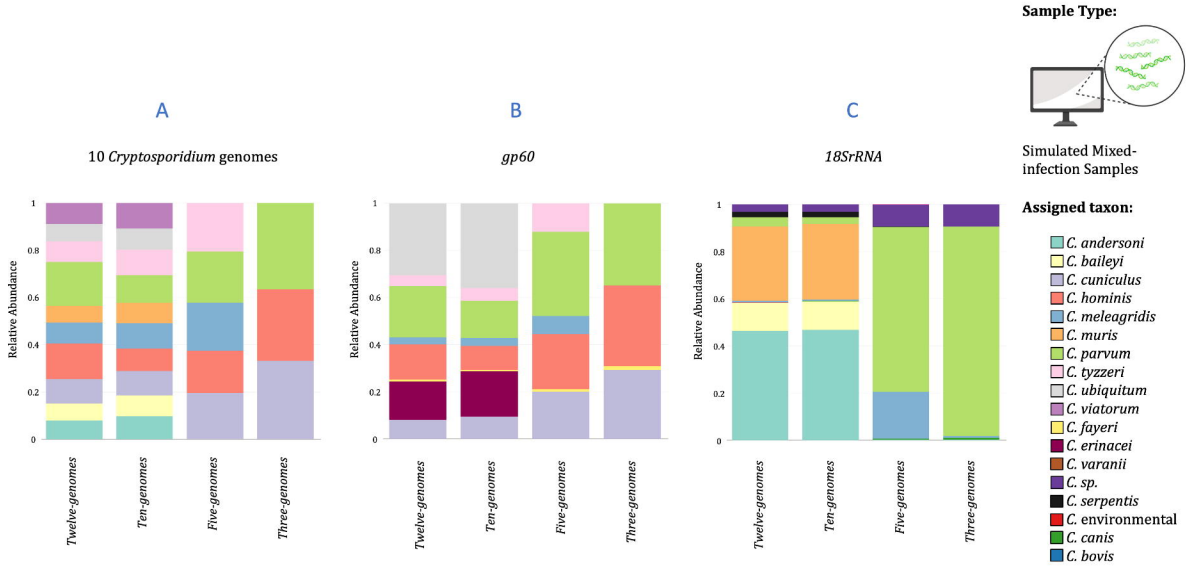
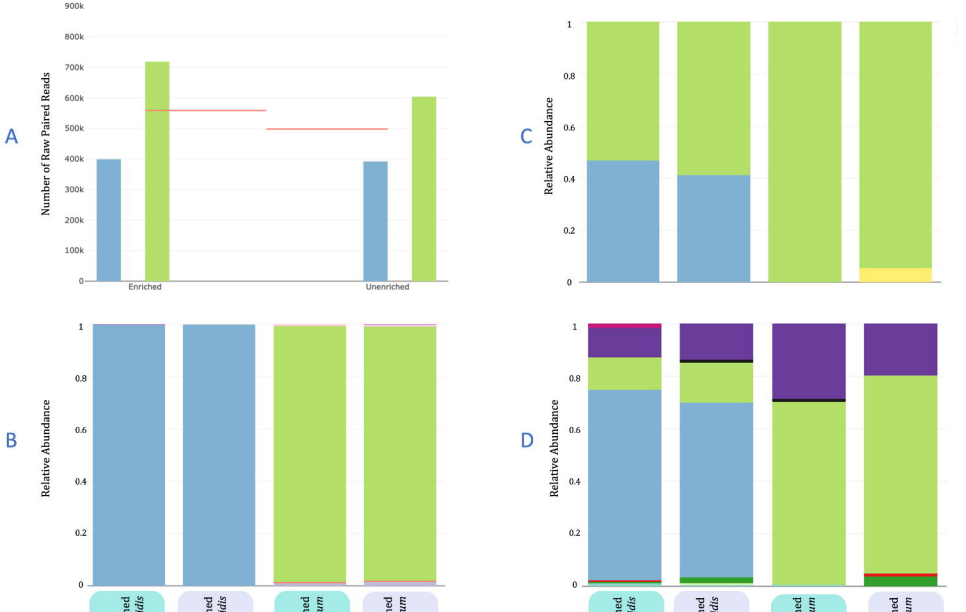


Figure 6. Summary statistics for sequence data from 100 clinical samples set prepared with the iNextEra protocol and iNext primers, comparing libraries enriched with CryptoCap_100k bait set versus unenriched libraries for validation. (A) Boxplot of percentage of reads mapped to *C. parvum* from double-enriched, single-enriched, and unenriched libraries, (B) Genome depth and breadth of mapped regions for each sample in double-enriched, single-enriched, and unenriched libraries normalized by 100,000 reads, (C) species assignment to each sample prepared with iNextEra protocol with nor further enrichment when mapped to a reference database containing ten *Cryptosporidium* genome sequences (10-CrypGS), (D) species assignment to each sample prepared with iNextEra protocol and single-enriched with CryptoCap_100K when mapped to a reference database containing ten *Cryptosporidium* genome sequences, (E) species assignment to each sample prepared with iNextEra protocol and double-enriched with CryptoCap_100K when mapped to a reference database containing ten *Cryptosporidium* genome sequences (10-CrypGS).


Reference:
10 Cryptosporidium genomes
gp60
18SrRNA
Sample Type:
Simulated Single-infection Samples
Assigned taxon:
C. andersoni
C. baileyi
C. cuniculus
C. hominis
C. meleagridis
C. muris
C. parvum
C. tyzzeri
C. ubiquitum
C. viatorum
C. fayeri
C. erinacei
C. varanii
C. sp.
C. serpentis
C. environmental
C. canis
C. bovis
A
B
C
D
E
F
Large Fragments
Small Fragments
Relative Abundance
0
0.2
0.4
0.6
0.8
1
C. andersoni
C. baileyi
C. cuniculus
C. hominis 30976
C. hominis UdeA01
C. meleagridis
C. muris
C. parvum
C. parvum ATCC
C. tyzzeri
C. ubiquitum
C. viatorum


A
10 Cryptosporidium genomes
B
gp60
C
18SrRNA
Relative Abundance
0
0.2
0.4
0.6
0.8
1
Twelve-genomes
Ten-genomes
Five-genomes
Three-genomes
Sample Type:
Simulated Mixed-infection Samples
Assigned taxon:
C. andersoni
C. baileyi
C. cuniculus
C. hominis
C. meleagridis
C. muris
C. parvum
C. tyzzeri
C. ubiquitum
C. viatorum
C. fayeri
C. erinacei
C. varanii
C. sp.
C. serpentis
C. environmental
C. canis
C. bovis



Sample Type:

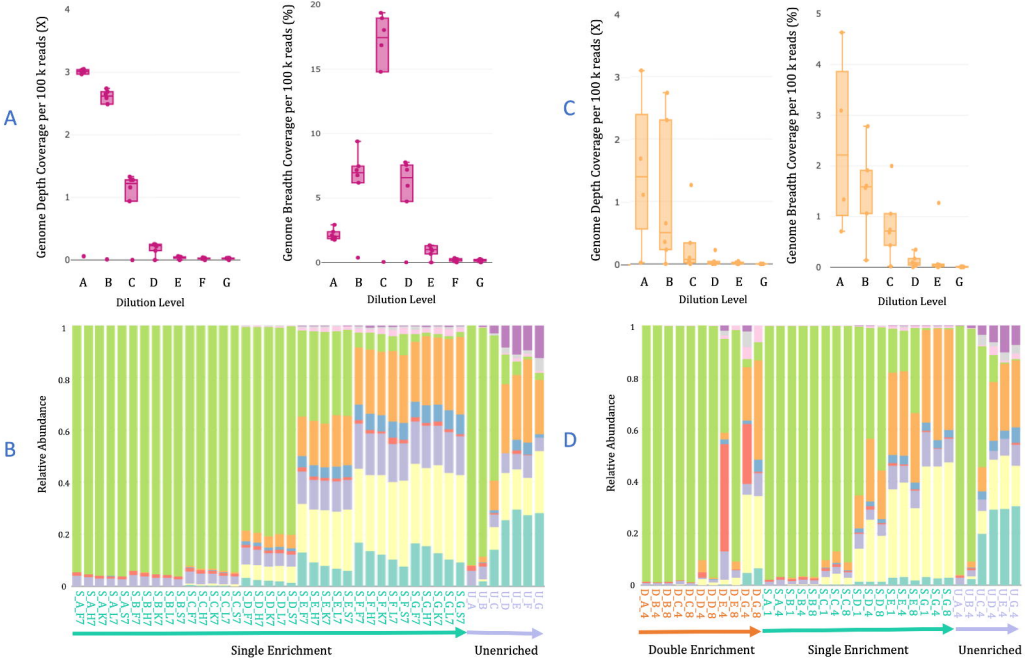
Mixed DNA Libraries from *C. parvum* and *C. meleagridis*

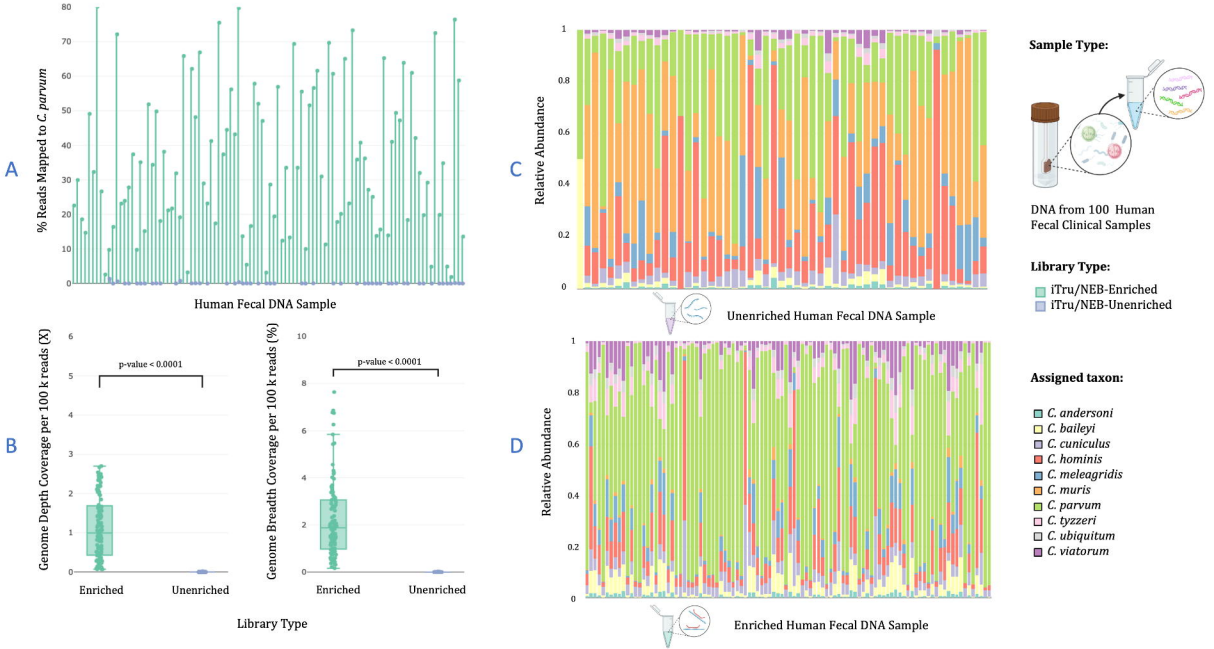
Library Type:

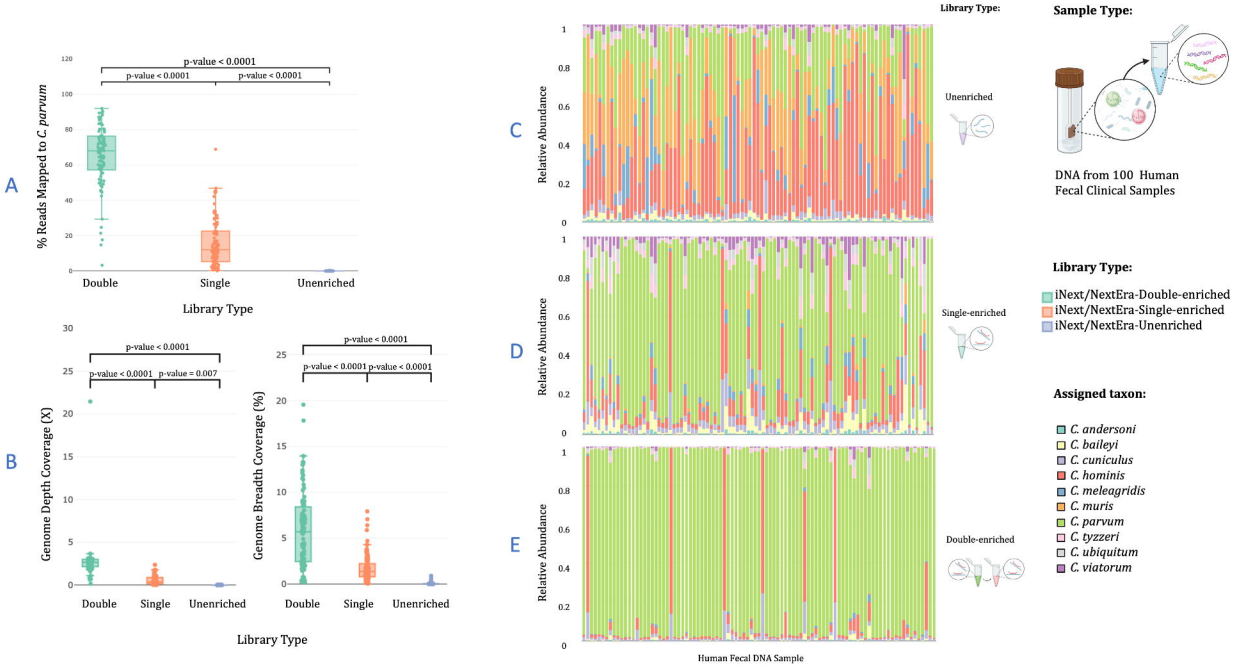
- iTru/NEB-Enriched
- iTru/NEB-Unenriched

Assigned taxon:

- *C. andersoni*
- *C. baileyi*
- *C. cuniculus*
- *C. hominis*
- *C. meleagridis*
- *C. muris*
- *C. parvum*
- *C. tyzzeri*
- *C. ubiquitum*
- *C. viatorum*
- *C. fayeri*
- *C. varanii*
- *C. sp.*
- *C. serpentis*
- *C.* environmental
- *C. canis*


A
Genome Depth Coverage per 100 k reads (X)
Genome Breadth Coverage per 100 k reads (%)
A B C D E F G
Dilution Level
C
B
D
Relative Abundance
Single Enrichment
Unenriched
Double Enrichment
Sample Type:
1:10
C. parvum DNA Samples Diluted on Calf Thymus DNA
Library Type:
iTru/NEB libraries
iNext/NextEra libraries
S: Single enrichment
D: Double enrichment
U: Unenriched
Assigned taxon:
C. andersoni
C. baileyi
C. cuniculus
C. hominis
C. meleagridis
C. muris
C. parvum
C. tyzzeri
C. ubiquitum
C. viatorum
Level of C. parvum Dilution
A: 1 ng
B: 0.1 ng
C: 0.01 ng
D: 0.001 ng
E: 0.0001 ng
F: 0.00001 ng
G: 0 ng
*Arrows indicate high input to low input of C. parvum


A
% Reads Mapped to C. parvum
Human Fecal DNA Sample
B
Genome Depth Coverage per 100 k reads (X)
p-value < 0.0001
Genome Breadth Coverage per 100 k reads (%)
p-value < 0.0001
Enriched
Unenriched
Library Type
C
Relative Abundance
Unenriched Human Fecal DNA Sample
D
Relative Abundance
Enriched Human Fecal DNA Sample
Sample Type:
DNA from 100 Human Fecal Clinical Samples
Library Type:
iTru/NEB-Enriched
iTru/NEB-Unenriched
Assigned taxon:
C. andersoni
C. baileyi
C. cuniculus
C. hominis
C. meleagridis
C. muris
C. parvum
C. tyzzeri
C. ubiquitum
C. viatorum


A
% Reads Mapped to C. parvum
p-value < 0.0001
p-value < 0.0001
p-value < 0.0001
Double
Single
Unenriched
Library Type
B
Genome Depth Coverage (X)
p-value < 0.0001
p-value < 0.0001
p-value = 0.007
Genome Breadth Coverage (%)
p-value < 0.0001
p-value < 0.0001
p-value < 0.0001
Library Type
C
D
E
Relative Abundance
Library Type:
Unenriched
Single-enriched
Double-enriched
Human Fecal DNA Sample
Sample Type:
DNA from 100 Human Fecal Clinical Samples
Library Type:
iNext/NextEra-Double-enriched
iNext/NextEra-Single-enriched
iNext/NextEra-Unenriched
Assigned taxon:
C. andersoni
C. baileyi
C. cuniculus
C. hominis
C. meleagridis
C. muris
C. parvum
C. tyzzeri
C. ubiquitum
C. viatorum