

Genome-targeted enrichment and sequencing of human-infecting *Cryptosporidium* spp.

Corresponding Author: Dr Natalia Bayona-Vásquez

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Introduction: The authors emphasize the value of CryptoCap_100k particularly for resolving mixed infections. How prevalent are mixed infections and why should these be resolvable by bait-enrichment methods but not conventional WGS upon oocyst purification? In addition, by mixed infections, do the authors mean specifically mixed infections with multiple *Cryptosporidium* species, or mixed infections with multiple genotypes of the same species, or both? Do they show data that tests that the CryptoCap_100k method improves on this? I would think the challenge for separating mixed infections and the potential to do this is dictated by sequencing depth, the sequencing technology (long vs short read) and post-sequencing informatics. The informatics would be largely the same for CES-seq vs standard short-read WGS, in which case mixed infections would be identified by allelic sequence heterogeneity. For suitably different species, these mixed types could be pulled apart by mapping to different references to identify species discrete SNPs (i.e., those found only in one and not the other species). But I do not know any specific way that closely related mixed genotypes could be pulled apart (other than through the same 'discrete' SNP approach) unless they could be reassembled using long-read sequencing, but that is not being explored here.

Introduction: I'm not sure about the argument that conventional methods require expensive deep sequencing. Can the authors clarify what they mean here? The cost of sequencing would be based on the required depth of coverage, which as far as I am aware would be the same for conventional WGS (assuming prior oocyst purification). There would added cost would be for the oocyst purification, but the only way I could see the sequencing component being more expensive is if the user attempted to undertake WGS without prior purification of the oocysts, and although I agree this would be much more expensive, as it would require much higher depth of sequencing, I don't think anyone would attempt this approach as it would be simply too impractical.

Introduction: In my opinion, the primary value of a target-enrichment based method for *Cryptosporidium* sequencing is that it would not require prior oocyst purification. This does reduce some costs and, more importantly, would increase sample throughput, but I think the main advantage would be that it provides a method that can be used for any sample, including a large library of archived samples, or from samples that are only available as total extracted DNA. I think based on the various comments above, it would be good if the authors considering streamlining their introduction and emphasizing a bit more clearly what they see the main impact of the method being.

Lines 145-161 - Did the authors simulate how mismatches between the bait probes and the target genome sequence might influence target-enrichment with the CryptoCap_100k among the species? If so, how was this done? For example, did the authors allow a certain number of mismatched SNPs per probe, and if so how many, and what was this threshold based on? From my reading of the methods, the in silico simulation does not make any consideration of mismatches (beyond perhaps a sequence alignment threshold - from the Supplementary Notes describing mapping of the simulated reads it seems only that the authors used default settings for bwa mem, is this correct and did the authors test whether these default settings were realistic for their purposes?) and appears to assume the baits will recognise each species more or less with the same efficiency and specificity. Am I misunderstanding? I see the authors undertake more detailed alignments of their simulated data using the 10-CrypGS database, but it's a little hard to determine how this relates to the primary analysis in this section.

Lines 204-213: Please provide an approximate estimate in oocyst equivalents for 1 through 0.00001 ng of Crypto DNA as this would give a rough context and perspective on the sensitivity as oocysts per gram of faeces. I see this is provided in the discussion, but I think it should be presented here along with the data showing how these oocyst / sporozoite equivalents were measured. Or are these values only calculated base on the assumption that one sporozoite contains 0.9 fg of DNA and 1 oocyst contains four sporozoites?

Line 215-218: I think its very important and am pleased to the see the authors have included breadth of coverage as a measure of sequencing performance in their analysis here, but it's not immediately clear to me what the mean by "a high percentage of breadth (14.7%)". Can they please clarify this. Is that 14.7% of the genome covered to a minimum depth (if so, what depth)? Similarly for lines 227-236 - breadth is an important measure and should be included, but please clarify how breadth was defined here.

Lines 227-238 - The authors mention a substantial reduction in the proportion of reads mapping post-enrichment with lower DNA concentrations. Did they explore what the other, non-Cryptosporidium reads are in these low DNA samples? Is this simply background DNA (e.g., human, bacterial, etc) coming through into the library because there is not enough template for the enrichment baits, or is this artefactual/chimeric sequences from the amplification phase post-enrichement? If the former, is it expected for background DNA to bleed through in this way? Could this be improved with additional washing steps post-enrichment but pre-library amplification? Alternatively, could the bait-enrichment step be applied post-library construction following amplification? If its chimeric/artifactual sequences, could these be further reduced with alternating the bead ratios in the clean up step, or reducing the number of post-enrichment amplification cycles (we have found this very helpful for low input samples in other enrichment or tiled amplicon based applications in my lab). Or again, could this be improved through additional clean-up or applying bait enrichment after library construction (would the bait probes allow this?). To be clear, I don't expect the authors to undertake any of these additional experiments for the current study, but maybe something to consider and if worthwhile / applicable, to mention briefly in discussion?

Lines 249-251: The increasing breadth of coverage with bait dilution is an interesting observation. Noting as the authors do that this difference is not statistically significant, its obviously not appropriate to speculate too much about what if anything this means, but in the context of my comments above, I wonder if again if the non-mapping sequences the authors find in low input samples tend to be more background sequences or chimeric artefacts and again whether some additional sensitivity could be squeezed out with varying (possibly slightly reducing) the post-enrichment amplification cycles, possibly in combination with reducing slightly the bait concentrations. Again, just something to consider. I don't expect the authors to undertake any testing of this in their study.

Lines 257-261: As above, please clarify how breadth of coverage is defined here.

Lines 264-272: Please provide a bit more detail on the clinical samples being tested here. In particular, do the authors have data on the number of oocysts shed per gram of faeces in these samples, or at least an estimated equivalent based on qPCR? Ideally it would be helpful to know if these patients were symptomatic. This provides important context for understanding the performance of the method in real-world samples. I.e., given that the control material shows that the overall performance of the method is dependent on having sufficient DNA in the sample to allow high coverage sequencing, are these clinical samples representative of typical infection intensities?

Lines 282-295: Similar question for the 100 human faecal samples used in this part of the study. I see that Table S17 provides qPCR Ct values, but these are only meaningful if they include comparable Ct values for a quantitative control series, ideally allowing the Ct values to be estimated as oocyst counts or oocysts per gram. This would also allow readers to understand if these samples are typical of the infection intensities one would expect to find in routine practice. As above, were any of these patients (and if so which ones) symptomatic for cryptosporidiosis at the time of sampling? Ideally, do the authors have corresponding oocyst per gram counts for the samples determined by the diagnostic clinical that collected the sample originally?

Lines 292-293 and Lines 326-328: Again, please define breadth of coverage here. Minimum 1x coverage? Or another cutoff?

Results: The authors have deployed the CryptoCap_100k baits to undertake bait-enriched sequencing of Cryptosporidium of ~100 faecal samples, but present only limited analysis of these data. I think this might relate to the low breadth of coverage the authors present (e.g., Lines 326-328). However, as noted above, I'm a bit confused about what the breadth statistics are showing. I think this needs to be clarified. If, largely speaking, the CryptoCap_100k method can only achieve coverage of ~5% of the genome (assuming that this what the breadth metric is showing). I've looked in the supplementary notes at how the authors have calculated breadth (i.e, number of bases achieving a minimum coverage depth) but I can't find what minimum coverage depth threshold was used (is it 1x? 5x? Something else?). I think it is important to resolve this as this impacts on the practical value of the method. In the current description of the data, the authors largely use the sequence data they have generated to determine the relative abundance of the Cryptosporidium species detected in the sample. However, if they are not achieving sufficient coverage to identify SNPs within the samples, or undertake population genetic analyses of the data etc, they really aren't producing data that is remarkably different from what could be generated with a suitably designed multiplex qPCR panel. The bait-enriched method is interesting, but I think really only if it is suitable to generate largely complete genomes from real-world samples at practical levels of sensitivity relevant to the typical infection intensities expected for individuals with cryptosporidiosis. If the authors do achieve enough coverage to undertake SNP typing of the genome (or at least for the predominant species within each sample), they should make use of these data to undertake a population genetic study of their sample set, thus illustrating the value of the method.

Results - Figure 5 and 6: How do the authors interpret the substantial apparent shift in assessed relative abundance of, for example, *C. hominis*, which predominates in unenriched samples (e.g., Fig 6C) and *C. parvum*, which predominates in the post-enriched samples (e.g, Fig 6d and 6e)? Does this reflect bias in the probes when used in real samples? I see that the authors have assessed this in their spiked control experiments (e.g., Fig 4), but that makes the results in Figs 5 and 6 that much more puzzling (unless I am misunderstanding). I think this needs to be assessed, ideally on a subset of samples using (at least) *C. hominis* vs *C. parvum* specific qPCRs / probes, in order to determine if the post-enriched species abundances are a true reflection of the species abundance in the samples. I see the authors mention in the discussion (line 451) that post-enrichment levels match previous testing of the samples by duplex real-time PCR, but I think this data needs to be matched to the samples in the supplementaries so that this can be evaluated and understood by the reader.

Results - Figure 5 and 6: How do the authors interpret the results showing that more or less all of their samples appear to contain sequence data representing numerous *Cryptosporidium* species? Are these mixed infections? This seems very unlikely, particularly noting that these figures appear to report reasonably high (~5-10%) relative abundance levels for species such as *C. baileyi* and *C. tyzzeri*, which are not typically considered to be human infectious. It seems more likely that these are reads that are being aligned incorrectly. Again, I think this needs to be tested independent, e.g., by species specific qPCR.

The authors should consider analysing raw sequence data from currently available *Cryptosporidium* WGS studies that have generated their data from purified oocysts. Most of these studies will have used either IMS beads to purify the samples or oocyst concentration of sucrose/ficoll density gradients. Unless there is evidence to suggest that these purifications are heavily biased for *C. parvum* / *C. hominis*, the raw sequence data should provide an accurate representation of the *Cryptosporidium* species in the samples. If these are analysed using the mapping to 10-CrypGS database, do the authors see the same complex mix of different *Cryptosporidium* species? If so, can they look more deeply at the reads being assigned to these different taxa to see if they contain discrete SNPs that, as far as is currently known, ought to be specific to each species (which would be stronger support for the common prevalence of complex species mixes consistent with their bait-enriched results).

Lastly, the authors could look at the reads from their bait-enriched data being assigned to the different species in their analysis. Where do these map? Do the authors see regions of the genome where they achieve high depth of coverage that contain mixes of reads mapping to the different species in their analysis? If so, if they look at the aligned reads in a genome viewer, can they see clear evidence of subpopulations of reads that contain mutations that are likely to be differentially diagnostic of the different species they are identifying? Or do the reads being attributed to the different species map to non-overlapping regions of the genome? If so, again, I suggest visualising inspecting the reads aligning to these regions to assess whether these alignments are likely correct. As this would seem to be more indicative of misaligned reads. The authors results may be correct, but they appear quite distinct from current understanding of the host-specificity of *Cryptosporidium* species and the prevalence of mixed infections and need robust testing to provide compelling evidence that this is indeed the case.

(Remarks on code availability)

The authors have provided a detailed account of their analytical steps including all major commands they used to run their analyses. These are well presented and appropriate and add significantly to the reproducibility of their study.

Reviewer #2

(Remarks to the Author)

The report describes the development of a new version of the *Cryptosporidium* genome capture assay (CryptoCap_100k) for the diagnosis of *Cryptosporidium* species and mixed infections. Its performance in capturing *Cryptosporidium* genomes is extensively evaluated using in silico analysis and mock samples. Its ability to assign *Cryptosporidium* species and detect mixed infections is further evaluated using 100 clinical samples. The new assay is shown to have a broader detection range than the previous assay (CryptoCap_75k), has high sensitivity, and allows the detection of multiple *Cryptosporidium* in individual samples. It also allows pooling of samples during analysis, reducing the costs associated with genomic sequencing of *Cryptosporidium* spp.

Main comments:

1. The title of the report ("Whole Genome Targeted Enrichment and Sequencing of Human-Infecting *Cryptosporidium* spp.") and much of the introduction on research needs is about whole genome sequencing (WGS) of *Cryptosporidium* spp. in clinical samples. Therefore, the focus of the study is expected to be on the performance of CryptoCap_100k in WGS analysis of *Cryptosporidium* spp. This is especially the case when CryptoCap_100k is built upon the existing CryptoCap_75k and genome capture is a technology widely used in WGS analysis of other pathogens. Therefore, the report should focus on the utility of CryptoCap_100k for *Cryptosporidium* sequencing, with less emphasis on the development and in silico evaluation of CryptoCap_100k itself.
2. The study could be much more vigorous with more data from experiments rather than in bioinformatics analysis and in silico simulations. Currently, two main tables (Tables 2 and 3) and two main figures (Fig. 1 and Fig. 2) from these analyses should be included in the supplementary materials. The associated descriptions of these data have taken up a large part of the Results section (L130-L172) and should be shortened. In addition, Fig. 3, Fig. 4 and most of the supplementary materials are also based on in silico analyses and mock samples. Only Fig. 5 and Fig. 6 present data generated from 100 clinical samples. Since these two figures present similar data generated from two library preparation protocols, only one should be presented as the main figure. Table 4 is another table comparing the data from the 100 samples processed with two library

preparation protocols. In addition, the concluding statement at the end of the abstract needs to be supported by data generated with real samples rather than in silico simulations.

3. There is a need for in-depth analysis of the data generated from the 100 clinical samples, especially the quality of the sequence data generated in relation to *Cryptosporidium* species and parasite load in the samples (as indicated by Ct values of qPCR analysis of the initial DNA rather than after library construction), such as percent genome coverage, average fold coverage, and genome assembly statistics. Such data would allow us to understand the performance of the genome capture and sequencing approach. Currently, these data are not presented in the Results section. I see buried in the Supplementary Results that for the 100 clinical samples analyzed, the depth and breadth of genome coverage were 28.1X and 31.3%, respectively. This result is presented in a very confusing way in the Results (L291-293): the depth and breadth of genome coverage were on average 1.12X and 2.25%, respectively, per 100,000 reads (for better understanding, Fig. 5B can be replaced with Fig. S35, while Fig. 6B can be replaced with Fig. S40). The reason for this disappointing result should be fully discussed.

4. The *Cryptosporidium* species identity of the initial 7 samples and the 100 clinical samples should be clearly stated, and samples with mixed species should be included. I could not find such information on the *Cryptosporidium*-positive samples anywhere in the main report and supplementary materials. Based on the data presented in Figure 5 and Figure 6, it appears that most of the 100 samples were from *C. parvum*, with a few from *C. hominis*. As CryptoCap_100k also targets four other species such as *C. meleagridis* and *C. viatorum*, samples positive for these should also be included. Agreement and disagreement between genotyping and CryptoCap_100k results, especially regarding mixed infections, should be fully described and discussed. In particular, the data presented in Figures 5 and 6 suggest that many of the clinical samples had moderate proportions of *C. baileyi*, *C. tyzzeri*, and other species for which there is limited evidence that they infect humans (as stated in L307-308). Has this finding been validated using species-specific PCR? There is no detailed discussion of the validity of this observation.

5. Since this report is about the development of CryptoCap_100k based on CryptoCap_75k, the comparison of the performance of the two in the detection of different species and mixed species should be done with real samples rather than in silico analysis. Again, it is difficult to accept that no samples of other targeted *Cryptosporidium* species were used in the evaluation of CryptoCap_100k.

Additional suggestions:

1. Some supplementary tables are nowhere to be found online, and all Excel tables online are not labeled. In addition, it is very difficult to find the cited supplementary tables and figures. There are just too many of them. Some of them should be merged or deleted. Most journals have limits on the number of supplementary figures. Some of the supplementary tables in this report cannot be found online, probably because the number of supplementary items in this manuscript already exceeds the limit. These problems make it very difficult to review the manuscript.

2. Abstract and elsewhere (L77-81...), the challenge of sequencing *Cryptosporidium* genomes in clinical samples is overstated. In the era of single-cell sequencing, protocols for sequencing single oocysts purified from clinical samples have been published with good genome coverage. Without considering cost, most *Cryptosporidium* genomes in human fecal samples can be sequenced by purification of oocysts by immunomagnetic separation, enrichment of extracted DNA by whole-genome amplification, and second- and third-generation sequencing. Considering the cost of fabricating the genome capture device, the greatest advantage of the genome capture approach is the ability to sequence *Cryptosporidium* genomes in obtained DNA from epidemiologic and outbreak investigations.

3. Introduction and L458-460. The problem of detecting mixed infections by PCR due to biased amplification also applies to CryptoCap_100k, as most capture baits are made from *C. parvum* sequences. If there is a mixed presence of *C. parvum* and *C. hominis* in a clinical sample, CryptoCap_100k would be expected to preferentially capture *C. parvum* sequences over *C. hominis* sequences. The failure to detect this in the computer simulation needs to be supported by vigorous testing with clinical specimens. Such statements should be toned down throughout the manuscript.

4. The differences in the baits between CryptoCap_100k and CryptoCap_75k should be clearly stated. If the *C. parvum* baits in CryptoCap_100k were all redesigned without consideration of what was included in CryptoCap_75k (this could be true based on the description in the Methods), this should be clearly stated in early parts of the report. Such information is necessary to understand the differences between the two capture systems and their performance.

5. The methods section is difficult to follow. It would be helpful to have a diagram of the procedures used for library construction, genome capture, and Illumina sequencing, including whether capture and amplification were performed in plates and sequencing was performed on pooled products. Many readers are not familiar with the genome capture process used. Currently, readers have to figure it out by carefully reading the entire Methods section. If whole genome capture was performed on metagenomic libraries generated from the extracted DNA instead of direct capture of sample DNA, the authors need to clearly state this early in the report.

6. The statement in L109-112 about including baits that capture subtelomeric regions of the genome in CryptoCap_100k is largely based on replacing IOWA-II with the better assembled IOWA-ATCC as the reference genome in the bait design (L378-379). While the IOWA-II genome is missing several telomeric ends, it has most of the subtelomeric sequences.

7. The length of the reads in statements in L285-286 should be provided. 150 or 250 PE?

8. The writing is acceptable, but the abstract reads poorly with some language errors.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Thank you to the authors for their thorough and considered responses to my comments and suggestions. These edits and responses have addressed each of my queries and I do not have further comments to add. My compliments to the authors on their high quality work. I am happy to recommend publication.

(Remarks on code availability)

The authors provide the dataset in a compressed file format and an R-Markdown script that allows all analyses and figures to be reproduced.

Reviewer #2

(Remarks to the Author)

The authors addressed some of the issues I raised. Therefore, the revised report is an improvement over the previous submission. However, because no new experimental data are presented, my other suggestions remain unresolved. Additionally, I still believe that some of the conclusions are too strong and not supported by the presented data. Results from in silico analyses, such as the data presented in Tables 1-3, should be moved to the supplementary materials.

Below are some specific suggestions, listed in the order I read them rather than in order of importance:

Abstract: The conclusions in the abstract should be modified. The data presented in this report do not support the conclusion that CryptoCap_100k "allows assessment of genetic variation via SNP calling." The relevant data presented in Figure 5 only support species identification. The nature of the clinical data presented, as well as the data presented in Figure 5, does not support identifying mixed species in the 91 clinical samples analyzed using the iNextEra library approach.

L37-40. It is general accepted that *C. tyzzeri* is not a human-infecting species, since there have been only a few reported cases of human infections. Please modify the statement and other related statements in the report.

L40-41, L56. Since spp. is a plural term, "Cryptosporidium spp. DNA" should be changed to "Cryptosporidium DNA".

Similarly, spp. should be followed by plural verbs.

L58. There are now far more than 39 recognized *Cryptosporidium* species.

L122. The word "data" is a plural term. Therefore, "This genomic data facilitates...." Should be replaced with "These genomic data facilitate...."

L133-136. This statement is difficult to understand. As it is written, the sentence implies that the comparison of the breadth of genome coverage between the two groups is made at different depths of coverage.

L134-135 and elsewhere. Keep one decimal place for percentage data.

L169-170. If the species assignment with 18S rRNA is only 70% correct, what is the rationale to include extra 18S rRNA baits in CryptoCap_100k?

L245-248. What are the implications for the observation of significant proportion of reads from many clinical samples to non-parvum species, including those with no evidence of human infectivity? Are these results from the NEB-iTru libraries reliable? This appears not the case with data from iNextEra libraries.

L272-278. The presence of both *C. hominis* and *C. cuniculus* reads in a few samples in Fig. 4H could be due to the high genetic identity between these two species instead of mixed infections.

L289. Data from which library construction approach (see above) are used in the Population Genomics Analyses? Logically, data from double enriched iNextEra libraries should be used in such analyses. Figure 5 suggests that indeed this may be the case but descriptions in this section fail to reflect this.

Discussion. This section should be significantly shorter. Currently, there is extensive repetition of results. Some statements about CryptoCap_100k's ability to perform SNP calling and comparative genomics should be modified. Similarly, the statements about detecting mixed species should be toned down. Since data from the double-enriched iNextEra library approach appears more reliable, this approach should be the focus of the discussion and recommended. Currently, the authors suggest that both library approaches and single and double enrichments are suitable for species assignment, mixed species detection, and comparative genomic analysis.

L357-364. These statements should be modified. Clearly, some species assignments from the NEB-iTru library approach are incorrect. As indicated above, the results shown in the study also do not support the conclusion that CryptoCap_100k sequencing enable SNP calling for population genomics.

Table 1 has no experimental data, therefore, should be moved to supplemental materials. Identity and overlap in the table should be defined. Without such information, it is difficult to understand the table.

Table 2. Genomes of *C. hominis* and *C. cuniculus* share greater than 98% sequence identity. It is unclear why the values of depth and breadth of genome coverages for *C. cuniculus* in the mapping of bait sequences are much smaller than those for *C. hominis*.

Tables 2 and 3. Table titles should have only partial sentence. Other descriptions should be moved to footnotes.

Figure 5. This figure has such low resolution that it is impossible to evaluate the accuracy of the associated descriptions in Results.

Figure S1. The Y-axis should be in the same scale to allow for more accurate comparison between species.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer #1 Remarks:

Introduction: The authors emphasize the value of CryptoCap_100k particularly for resolving mixed infections. How prevalent are mixed infections and why should these be resolvable by bait-enrichment methods but not conventional WGS upon oocyst purification? In addition, by mixed infections, do the authors mean specifically mixed infections with multiple *Cryptosporidium* species, or mixed infections with multiple genotypes of the same species, or both?

These are excellent points made by the reviewer. We believe that the core value of CryptoCap_100K is to generate genomic data from clinical samples where the concentrations of target DNA are low, without the need for oocyst purification. Thus, we have restructured our introduction to address this goal first before introducing mixed infections. Still, there are extended benefits from generating this type of data, which include an approximation for studying mixed infections, which we now discuss more carefully.

- We have included information on the introduction, which describes mixed infections and their prevalence in lines 86–96, highlighting the importance of studying them.
- Regarding the question of why bait-enrichment methods should resolve mixed infections of clinical samples, but not conventional WGS upon oocyst purification.
 - We explain in the next question below how we demonstrate in this study that our bait-enrichment method is useful, among other analyses, for parsing mixed infection samples. Also, we have included a full Discussion section on this regard in lines: 454–472, under the section “Crypto_Cap_100K Enables the Detection of Mixed Infections”

- We believe both genomic methods, CryptoCap_100K and WGS, can help answer similar questions. Introduction to approaches that use WGS of *Cryptosporidium* and their limitations are now found in lines 75–85.
- Regarding the reviewer’s question, do we mean by 'mixed infections' infections with multiple *Cryptosporidium* species, multiple genotypes of the same species, or both?
 - We have made several edits to ensure greater care in these descriptions throughout the introduction, results, and discussion. For example, when possible, we refer to “mixed-species infections” or highlight “mixed infections with strains of the same species”. Importantly, we were able to assess and identify both types with our data.

In summary, we thank the reviewer for pointing out the lack of background and specificity in defining mixed infections and in describing the benefits of our capture system. We hope that some of the edits considered throughout the manuscript help resolve this issue. In summary, we believe that CryptoCap_100K is a valid alternative to methods that avoid oocyst purification, such as whole-genome amplification, while still achieving a high success rate that allows for the performance of mixed infection analyses.

Do they show data that tests that the CryptoCap_100k method improves on this? I would think the challenge for separating mixed infections and the potential to do this is dictated by sequencing depth, the sequencing technology (long vs short read) and post-sequencing informatics. The informatics would be largely the same for CES-seq vs standard short-read WGS, in which case mixed infections would be identified by allelic sequence heterogeneity. For suitably different species, these mixed types could be pulled apart by mapping to different references to identify species discrete SNPs (i.e., those found only in one and not the other species). But I do not know any specific way that closely related mixed genotypes could be pulled apart (other than through the same 'discrete' SNP approach) unless they could be reassembled using long-read sequencing, but that is not being explored here.

Thank you for bringing this comment to our attention, as it prompted us to perform analyses that directly help parse apart mixed infections (species or strains within a species) in haploid systems such as

Cryptosporidium, in addition to those *in silico* and *in vitro* analyses we presented in our previous version, still included.

To directly answer the question from the reviewer: “Do they show data that tests whether the CryptoCap_100k method improves on this [resolving mixed infections]?” In this paper, we present several types of evidence to demonstrate how our approach enables the parsing of mixed infections by:

- a. *In silico* simulations demonstrated that the bait diversity considered in this study enables the capture of a wide range of *Cryptosporidium* species and simulated mixed-species infections provided evidence that our method accurately detects different species in a mix (Fig. S13).
- b. *In vitro* mock communities in the lab also provided evidence that enrichment with CryptoCap_100K enabled the detection and discerning between species in a mock community of *C. parvum* and *C. meleagridis*. However, we found slight biases towards *C. parvum* in terms of relative abundance (Fig. 1B).
- c. Analyses of species assignment suggested the likely presence of mixed-species infection samples in our clinical sets (Figs. 4H and S32), which were corroborated in the only sample from this pool that was included in the population genomics and deconvolution analyses (Fig. 5A).
- d. New analyses of clinical samples include those for deconvoluting genetic data to detect mixed infection samples using software developed for haploid organisms such as *Cryptosporidium*. This software allows for the detection of allele imbalances that translate to multiple haplotypes (either with multiple species or strains from one species). Double- and single-enriched libraries suggest that sample UKH149 is likely a mixed-species sample (Fig. 5A). Heterozygosity measurements also support these results. Importantly, we also find evidence of samples with likely two strains of *C. parvum* (UK347 and UK391, Fig. 5A, Table S18–S19)
- e. Overall, we provide evidence on how our method can be used to understand if samples contain mixed infections of different natures. We expand on reported methods for detecting mixed infections and their challenges in lines: 454–472.

Regarding the challenges that the reviewer mentions — sequencing depth, short reads, and bioinformatic analyses — we and other studies have demonstrated that short-read sequencing at good genome coverage allows for the untangling of mixed infections. In this paper, sequencing reads, and their variation allowed

the discrimination of species (Figs. 4H and 5A), and SNP calling of double-enriched libraries allowed for the detection of multiple haplotypes within samples (Fig. 5A). We provide one pipeline to assess if multiple species are present in a sample (Figs. 3F, 4G-4H; Supplementary Notes # 12.2.1.2 and 12.2.2.2), and a second independent pipeline to assess samples containing any mixed infection (Fig. S87; Supplementary Notes # 13.6.1). This was possible thanks to the use of tools developed for other haploid parasites, such as DEploid, which offers the option of not using a reference panel and was designed to detect mixed infections involving multiple species or strains.

Still, we agree with the reviewer that challenges remain in bioinformatic analyses, in part due to the lack of comprehensive databases with complete genomes that accurately reflect the diversity of and within species to compare samples against and assign proper strains and species. We hope that as more clinical samples get surveyed with methods such as CryptoCap_100k, such databases and reference panels can be generated and populated, as suggested in lines 484–486.

Additionally, as the reviewer suggested, we explored heterogeneity within a sample by estimating heterozygosity levels for what is a haploid parasite, using this measure to assess multiple strains in a sample (Figs. S81-S86). Still, we recognize that we must exercise caution regarding our method, given that library preparation can introduce biases in the proportions of species detected, particularly towards *C. parvum*, as also mentioned in our discussion (lines 464–466).

The suggestion from the reviewer that mixed infections can be detected by mapping to different references to identify species-discrete SNPs (i.e., those found only in one species and not the other) is on the right track. We want to highlight that this was similar to our approach in the *in silico* and experimental analyses, called “species assignment” throughout our paper. Sequence reads from *in vitro* and clinical samples were mapped to 10_CrypGS (one fasta file containing reference genomes from 10 *Cryptosporidium* species), and the number of reads that mapped to each of the ten reference genomes was counted and plotted (Figs. 1B, 2B–2D, S32, 3E–3F, 4F–H). We did not have the need to dissect species-specific SNPs to accurately identify the species present in the dataset; in this case, mismatches originating from those SNPs were reflected in mapping quality and quantity. Because detecting discriminant SNPs for all human-infecting *Cryptosporidium* species was not the focus of our study, we don’t expand on this.

Finally, the reviewer mentions that they are not familiar with the methods by which closely related mixed genotypes can be separated and that these methods were not explored in the paper. This was correct on our previous submission, given that the focus was on the capture system rather than on analysis methods for the field. However, in this new version, we now incorporate the DEploid method (Zhu et al., 2019) for strain deconvolution from genome sequencing data. Please refer to *Population Genomics Analyses of Clinical Samples* in the Results section, specifically lines 316–327, and to *CryptoCap_100K Enables the Detection of Mixed Infections* in the Discussion section. This method has been validated for other parasites; thus, although useful, long-sequencing reads are not necessary for deconvolving haplotypes, and information on genotype and allelic depth from short-read data is useful to detect allele imbalances within samples.

Introduction: I'm not sure about the argument that conventional methods require expensive deep sequencing. Can the authors clarify what they mean here? The cost of sequencing would be based on the required depth of coverage, which as far as I am aware would be the same for conventional WGS (assuming prior oocyst purification). There would added cost would be for the oocyst purification, but the only way I could see the sequencing component being more expensive is if the user attempted to undertake WGS without prior purification of the oocysts, and although I agree this would be much more expensive, as it would require much higher depth of sequencing, I don't think anyone would attempt this approach as it would be simply too impractical.

The reviewer is on the right track. We meant that sequencing costs are higher for samples that have not been oocyst-purified or enriched for *Cryptosporidium*. We rephrased the paragraph (lines 63–85), hoping to enhance clarity. Direct sequencing of fecal samples has been used by some authors, such as Hadfield et al. (2015), and thus is not impractical. However, these methods rely on a sufficient sample volume and an adequate concentration of oocysts, and achieve WGS data for 53% of the samples, indicating a moderate success rate. This information is now included in the lines pointed out above.

Introduction: In my opinion, the primary value of a target-enrichment based method for *Cryptosporidium* sequencing is that it would not require prior oocyst purification. This does reduce some costs and, more

importantly, would increase sample throughput, but I think the main advantage would be that it provides a method that can be used for any sample, including a large library of archived samples, or from samples that are only available as total extracted DNA. I think based on the various comments above, it would be good if the authors considering streamlining their introduction and emphasizing a bit more clearly what they see the main impact of the method being.

Thanks to the reviewer for this comment! We agree with the primary value stated by the reviewer. We hope that the edits made to our paper accurately reflect the project's goal. Among many other edits, to ensure the message resonates with the audience, we have streamlined the discussion to incorporate some of the language suggested by the reviewer, as seen on lines 352–356.

Lines 145-161 - Did the authors simulate how mismatches between the bait probes and the target genome sequence might influence target-enrichment with the CryptoCap_100k among the species? If so, how was this done? For example, did the authors allow a certain number of mismatched SNPs per probe, and if so how many, and what was this threshold based on? From my reading of the methods, the *in silico* simulation does not make any consideration of mismatches (beyond perhaps a sequence alignment threshold - from the Supplementary Notes describing mapping of the simulated reads it seems only that the authors used default settings for bwa mem, is this correct and did the authors test whether these default settings were realistic for their purposes?) and appears to assume the baits will recognise each species more or less with the same efficiency and specificity. Am I misunderstanding? I see the authors undertake more detailed alignments of their simulated data using the 10-CrypGS database, but it's a little hard to determine how this relates to the primary analysis in this section.

Thanks to the reviewer for this comment. Our *in silico* simulations did not assess how mismatches between bait probes and target genomes may influence enrichment. We used default parameters from BWA, which uses a k-value to determine the maximum number of mismatches given the read length of the query (in this case, the baits). Li and Durbin (2009) describe it as:

“Given a read of length m , BWA only tolerates a hit with at most k differences (mismatches or gaps), where k is chosen such that $<4\%$ of m -long reads with 2% uniform base error rate may

contain differences more than k. With this configuration, for 15–37 bp reads, k equals 2; for 38–63 bp, k=3; for 64–92 bp, k=4; for 93–123 bp, k=5; and for 124–156 bp reads, k=6.”

Because our baits are 120 nt long, the number of mismatches that BWA allows is for K=5, and we did not test different settings. We hope this clarifies the questions from the reviewer and explains why the *in-silico* analyses in our paper are a first approach to validate whether we could obtain meaningful data before synthesis and experimental design. We introduced this reasoning in line 522 and also have included the information from Li and Durbin (2009) in Supplementary Notes #3.

But yes! It is important, as stated in the Supplementary Notes and now in the main manuscript, to highlight that our *in silico* methods assume an error-free and bias-free enrichment process. This wording suggests that our simulation settings were not necessarily realistic, as they assume that variables such as the identity of the probe to the genome and the thermodynamics of the specific reaction, which affect hybridization and capture, do not vary significantly. The goal of the simulations was to create a simple scenario as proof of concept that our method could capture a wide diversity of *Cryptosporidium* species before synthesizing the bait set. Since the outcome analyses reflected the input, we felt confident in proceeding with *in vitro* tests and on real samples. We believe that *in silico* and *in vitro* tests are a valuable complement, having validated the efficiency and specificity of our bait set. Besides these goals, *in silico* analyses also provided an understanding of the specificity of our prob set, when baits were mapped to more than 1,000 genome sequences.

Finally, regarding the last comment, the reviewer states that “... it is a little hard to determine how this [more detailed alignments of their simulated data using the 10-CrypGS database] relates to the primary analysis in this section.” We want to highlight that the *in silico* analysis of mapping reads to 10-CrypGS and other databases was also used in other experiments, detailed in the Methods and Supplementary Notes.

The reviewers will notice that, thanks to suggestions from the second reviewer, we have relocated a substantial portion of the *in silico* simulations from the main manuscript, including figures and tables, to the Supplementary Results, Supplementary Tables, and Supplementary Figures so that we can focus on experimental and clinical data in the main text, and thus on the applicability of our method.

We have also been more intentional about comparing the multiple methods from *in silico* to real-world samples in the Discussion (e.g., lines 374–376).

Lines 204-213: Please provide an approximate estimate in oocyst equivalents for 1 through 0.00001 ng of Crypto DNA as this would give a rough context and perspective on the sensitivity as oocysts per gram of faeces. I see this is provided in the discussion, but I think it should be presented here along with the data showing how these oocyst / sporozoite equivalents were measured. Or are these values only calculated base on the assumption that one sporozoite contains 0.9 fg of DNA and 1 oocyst contains four sporozoites?

We appreciate the suggestion from the reviewer. However, because these lines belong to the “Results” section of our manuscript, and because we tweaked experimentally the amount of DNA going into the reaction, which we purchased from commercial vendors (DNA, not oocysts), and we did not tweak directly the number of oocysts used for DNA extraction. We believe our current language more accurately reflects the experiment and does not mislead the reader into thinking we tested the amount of DNA obtained from different numbers of oocysts. However, as the reviewer points out, we included the estimates in our abstract, introduction, and discussion of oocyst equivalents. In these sections, we estimated the values using a formula that yields ~9 fg of DNA per sporozoite and ~40 fg per oocyst. Please see lines 382–394.

Line 215-218: I think its very important and am pleased to the see the authors have included breadth of coverage as a measure of sequencing performance in their analysis here, but it's not immediately clear to me what the mean by "a high percentage of breadth (14.7%)". Can they please clarify this. Is that 14.7% of the genome covered to a minimum depth (if so, what depth)? Similarly for lines 227-236 - breadth is an important measure and should be included, but please clarify how breadth was defined here.

The definition of breadth and how it was estimated can be found in lines 514–516 and Supplementary Notes:

The breadth of coverage, as reported throughout this manuscript, was determined by estimating the number of bases covered in the alignment over the reference length, with a minimum coverage depth of 1X.

In this specific section of the manuscript, the text read in the previous version: "... significant differences between 0.01 ng, which has a high percentage of breadth (14.7%), and all other comparisons...". The second reviewer brought up the question of why this value was considered high. This sentence referred to normalized information by 100,000 sequencing reads. Given the spread confusion from normalized values, we have moved all normalized values and figures to supplements and included raw depth and breadth estimates in the main manuscript. Please see lines 194–196.

Lines 227-238 - The authors mention a substantial reduction in the proportion of reads mapping post-enrichment with lower DNA concentrations.

Thanks to the reviewer for the following questions!

- Did they explore what the other, non-Cryptosporidium reads are in these low DNA samples? In the previous version, we did not investigate the origin of non-target reads in the dilution experiment. Now we do!
- Is this simply background DNA (e.g., human, bacterial, etc) coming through into the library because there is not enough template for the enrichment baits, or is this artefactual/chimeric sequences from the amplification phase post-enrichment? Following the reviewer's suggestion, we present methods, results, and discussion of the assessment of non-target reads to determine if they originated from the background community composed of *Bos taurus* DNA. Please refer to Fig. S22, where we found that indeed these correspond to such.
- If the former, is it expected for background DNA to bleed through in this way? Yes, off-target DNA is expected to some extent in capture experiments. For context, other studies that have developed baits with myBaits (company used in this study), such as Hoffberg et al. (2016) in plants, Beaudry et al. (2021) for bacterial 16SrRNA, and Beaudry et al. (2021) for bacterial resistomes, found on average 60%–90% of target reads in purified samples and 45%–68% in

environmental samples. Simon and Machyna (2019) have addressed the potential sources of background DNA after enrichment, which is now discussed in lines 391–392.

- Could this be improved with additional washing steps post-enrichment but pre-library amplification? Alternatively, could the bait-enrichment step be applied post-library construction following amplification? Background DNA can be diminished through post-enrichment stringent washing steps, but this also depletes the amount of target DNA in the reaction. This tradeoff can be detrimental for samples that need it, such as those with low target DNA input. We demonstrate in our paper that we obtain genomic data for SNP calling in up to 85% of clinical samples, despite wide variations in qPCR C_T values. Therefore, with this protocol being successful, we don't explore additional washes. The enrichment occurs post-library construction, following amplification, now depicted in Fig. 6.

- If its chimeric/artifactual sequences, could these be further reduced with alternating the bead ratios in the clean up step, or reducing the number of post-enrichment amplification cycles (we have found this very helpful for low input samples in other enrichment or tiled amplicon based applications in my lab). Or again, could this be improved through additional clean-up or applying bait enrichment after library construction (would the bait probes allow this?). To be clear, I don't expect the authors to undertake any of these additional experiments for the current study, but maybe something to consider and if worthwhile / applicable, to mention briefly in discussion?

In designing our capture experiment, we considered benchmark guidelines to prevent the formation of PCR chimeras, which include maintaining high PCR annealing temperatures (our post-enrichment PCRs are performed at 60°C and limiting the number of cycles to 16 or fewer to prevent PCR from reaching a plateau. We also don't see any evidence for chimera formation in PCR pools in our data analyses.

Lines 249-251: The increasing breadth of coverage with bait dilution is an interesting observation. Noting as the authors do that this difference is not statistically significant, its obviously not appropriate to speculate too much about what if anything this means, but in the context of my comments above, I wonder if again if the non-mapping sequences the authors find in low input samples tend to be more background sequences or chimeric artefacts and again whether some additional sensitivity could be

squeezed out with varying (possibly slightly reducing) the post-enrichment amplification cycles, possibly in combination with reducing slightly the bait concentrations. Again, just something to consider. I don't expect the authors to undertake any testing of this in their study.

Following our previous response, we expect non-mapping sequences to be the non-target DNA present in the community, in this case, *Bos taurus* (Fig. S22).

Lines 257-261: As above, please clarify how breadth of coverage is defined here.

Same definition as above, also included in the Materials and Methods, lines 514–516.

Lines 264-272: Please provide a bit more detail on the clinical samples being tested here. In particular, do the authors have data on the number of oocysts shed per gram of faeces in these samples, or at least an estimated equivalent based on qPCR?

These are great points! Unfortunately, we don't have information on the number of oocysts shed per gram of feces in our clinical samples.

Ideally it would be helpful to know if these patients were symptomatic. This provides important context for understanding the performance of the method in real-world samples. I.e., given that the control material shows that the overall performance of the method is dependent on having sufficient DNA in the sample to allow high coverage sequencing, are these clinical samples representative of typical infection intensities?

Unfortunately, we don't have such information. We agree with the reviewer that, when available, these analyses should include the number of oocysts per sample, along with various metadata, such as whether the individual is symptomatic, the transmission mode, and whether the infection is recurring. The reviewer's suggestions align with the future directions we propose in the closing of our discussion, lines 482–486.

Lines 282-295: Similar question for the 100 human faecal samples used in this part of the study. I see that Table S17 provides qPCR Ct values, but these are only meaningful if they include comparable Ct values for a quantitative control series, ideally allowing the Ct values to be estimated as oocyst counts or oocysts per gram. This would also allow readers to understand if these samples are typical of the infection intensities one would expect to find in routine practice. As above, were any of these patients (and if so which ones) symptomatic for cryptosporidiosis at the time of sampling? Ideally, do the authors have corresponding oocyst per gram counts for the samples determined by the diagnostic clinical that collected the sample originally?

Unfortunately, we don't have qPCR C_T values for a quantitative control series. Also, using qPCR values to estimate the number of *Cryptosporidium* oocysts in a sample has a great variation in accuracy, especially at low concentrations or in complex samples. Although we don't know the number of oocysts, the overall abundance of oocysts in stool samples tends to be low in infected patients, and the ones studied here are likely no exception. However, we discussed the relationship of C_T values and retrieval of target genomic data with and without CryptoCap_100k in lines 409–418. We demonstrate that our method retrieves informative genomic-wide data for ~85% of fecal samples with average C_T values of 25.4.

The set of 100 human clinical samples was obtained from Dr. Rachel Chalmers' laboratory at the *Cryptosporidium* Reference Unit of Public Health Wales, as part of their molecular surveillance program. All are representative of *Cryptosporidium* cases, mainly in young children and adults of childbearing age. However, we don't have the metadata with patients' symptomatology. Still, we believe that our samples represent typical infections, which highlights the value of our study.

Lines 292-293 and Lines 326-328: Again, please define breadth of coverage here. Minimum 1x coverage? Or another cutoff?

Same definition as above throughout the full paper.

Results: The authors have deployed the CryptoCap_100k baits to undertake bait-enriched sequencing of *Cryptosporidium* of ~100 faecal samples, but present only limited analysis of these data. I think this might

relate to the low breadth of coverage the authors present (e.g., Lines 326-328). However, as noted above, I'm a bit confused about what the breadth statistics are showing. I think this needs to be clarified. If, largely speaking, the CryptoCap_100k method can only achieve coverage of ~5% of the genome (assuming that this is what the breadth metric is showing). I've looked in the supplementary notes at how the authors have calculated breadth (i.e., number of bases achieving a minimum coverage depth) but I can't find what minimum coverage depth threshold was used (is it 1x? 5x? Something else?). I think it is important to resolve this as this impacts on the practical value of the method. In the current description of the data, the authors largely use the sequence data they have generated to determine the relative abundance of the *Cryptosporidium* species detected in the sample. However, if they are not achieving sufficient coverage to identify SNPs within the samples, or undertake population genetic analyses of the data etc, they really aren't producing data that is remarkably different from what could be generated with a suitably designed multiplex qPCR panel. The bait-enriched method is interesting, but I think really only if it is suitable to generate largely complete genomes from real-world samples at practical levels of sensitivity relevant to the typical infection intensities expected for individuals with cryptosporidiosis. If the authors do achieve enough coverage to undertake SNP typing of the genome (or at least for the predominant species within each sample), they should make use of these data to undertake a population genetic study of their sample set, thus illustrating the value of the method.

We hope our answers and edits have clarified how we use and estimate the breadth of coverage, as well as what we mean when we mention normalized values. We found that in clinical samples, our method can achieve, on average, 50% genome coverage at a depth of 1X (Table S15). The 5% of genome coverage at depth 1X is per 100,000 sequencing reads, which is a very, very, very low sequencing coverage compared to the average of 4 to 5 million sequence reads obtained per sample. We used the normalized measure to make comparisons among samples, which, thanks to suggestions of the second reviewer, have now been moved to supplementary information to avoid confusion.

We initially did not include SNP calling and diversity analyses. However, we are very grateful for the feedback of the reviewer, as it led us to incorporate population genomic analyses from thousands of SNP calls that validate our previous analyses and solidify the use of our CryptoCap_100k. Please see the methods, results, and discussion on this regard.

Results - Figure 5 and 6: How do the authors interpret the substantial apparent shift in assessed relative abundance of, for example, *C. hominis*, which predominates in unenriched samples (e.g., Fig 6C) and *C. parvum*, which predominates in the post-enriched samples (e.g, Fig 6d and 6e)?

Here, it is important to consider that for clinical samples, 0.02–0.15% of sequencing reads mapped (e.g., lines 217 and 232). These very few reads mapped to genome locations from multiple species, including non-human infecting species. In lines 220–221 and 244–245, when discussing patterns in unenriched clinical samples, we now mention: “Species assignments for unenriched libraries resemble the complex background from LOD experiments.” Thus, we conclude that unenriched samples are insufficient to assess species composition and are likely low-quality mapping reads. This is validated by the fact that when using such mapped reads to call SNPs, we get high levels of missing data (Supplementary Results) with no discernible biological pattern, as those obtained for enriched samples (e.g., lines 314–315).

Does this reflect bias in the probes when used in real samples? I see that the authors have assessed this in their spiked control experiments (e.g., Fig 4), but that makes the results in Figs 5 and 6 that much more puzzling (unless I am misunderstanding). I think this needs to be assessed, ideally on a subset of samples using (at least) *C. hominis* vs *C. parvum* specific qPCRs / probes, in order to determine if the post-enriched species abundances are a true reflection of the species abundance in the samples. I see the authors mention in the discussion (line 451) that post-enrichment levels match previous testing of the samples by duplex real-time PCR, but I think this data needs to be matched to the samples in the supplementaries so that this can be evaluated and understood by the reader.

Thanks to the fact that these samples had previously undergone multi-locus typing (not from this study), as indicated by the UKP (*C. parvum*) and UKH (*C. hominis*) sample names, we skipped the qPCR step and matched the data to our results. In the results and discussion, we acknowledge the possibility of biases due to the number of baits used for *C. parvum*. We discuss closely the most obvious sample, UKH149 (e.g., lines 462–466), which was initially typed as *C. hominis*. However, it appears as a mixed-species infection sample of *C. hominis* and *C. parvum*, with higher proportions of *C. parvum* after double enrichment compared to single enrichment.

Results - Figure 5 and 6: How do the authors interpret the results showing that more or less all of their samples appear to contain sequence data representing numerous *Cryptosporidium* species? Are these mixed infections? This seems very unlikely, particularly noting that these figures appear to report reasonably high (~5-10%) relative abundance levels for species such as *C. baileyi* and *C. tyzzeri*, which are not typically considered to be human infectious. It seems more likely that these are reads that are being aligned incorrectly. Again, I think this needs to be tested independent, e.g., by species specific qPCR.

The results presented in Figures 5 and 6, now 3 and 4, that suggested that clinical samples may have *C. baileyi*, *C. tyzzeri*, and other species with limited evidence that they infect humans, may be the outcome of, as mentioned above, very low mapping quality of the very few hits and/or contamination of the genome assemblies of species such as *C. muris* with likely bacterial reads.

The authors should consider analysing raw sequence data from currently available *Cryptosporidium* WGS studies that have generated their data from purified oocysts. Most of these studies will have used either IMS beads to purify the samples or oocyst concentration of sucrose/ficoll density gradients. Unless there is evidence to suggest that these purifications are heavily biased for *C. parvum* / *C. hominis*, the raw sequence data should provide an accurate representation of the *Cryptosporidium* species in the samples. If these are analysed using the mapping to 10-CrypGS database, do the authors see the same complex mix of different *Cryptosporidium* species? If so, can they look more deeply at the reads being assigned to these different taxa to see if they contain discrete SNPs that, as far as is currently known, ought to be specific to each species (which would be stronger support for the common prevalence of complex species mixes consistent with their bait-enriched results).

It is not our goal to introduce mapping to 10-CrypGS as a gold standard for species assignment; rather, it was introduced as an exploration of the potential presence of species in a sample. In this new version, we have included more robust analyses, including SNP calling, which is the gold standard for phylogenetic and population genomic studies. We hope these new analyses align more closely with the reviewer's suggestion to check for SNPs that could be specific to species.

Lastly, the authors could look at the reads from their bait-enriched data being assigned to the different species in their analysis. Where do these map? Do the authors see regions of the genome where they achieve high depth of coverage that contain mixes of reads mapping to the different species in their analysis? If so, if they look at the aligned reads in a genome viewer, can they see clear evidence of subpopulations of reads that contain mutations that are likely to be differentially diagnostic of the different species they are identifying? Or do the reads being attributed to the different species map to non-overlapping regions of the genome? If so, again, I suggest visualising inspecting the reads aligning to these regions to assess whether these alignments are likely correct. As this would seem to be more indicative of misaligned reads. The authors results may be correct, but they appear quite distinct from current understanding of the host-specificity of *Cryptosporidium* species and the prevalence of mixed infections and need robust testing to provide compelling evidence that this is indeed the case.

Thanks to this reviewer's suggestion, we have produced Fig. S40 (<https://doi.org/10.6084/m9.figshare.29621024.v1>), which illustrates the depth and breadth of mapping against Cryp-10GS. It is notable that for unenriched samples, many reads map to all genomes, which introduces a bias towards *C. baileyi* and *C. muris*. Now, when we focus on enriched libraries, we observe that reads map preferentially to the genomes to which they were assigned in the multi-locus typing system and our species assignment analyses.

The authors have provided a detailed account of their analytical steps including all major commands they used to run their analyses. These are well presented and appropriate and add significantly to the reproducibility of their study.

Thanks to this reviewer's compliment. As we organized our supplemental materials in this new version to enhance readability, we maintained the same rigor in providing all necessary information to ensure the reproducibility of this study, including code for bioinformatic and downstream data analyses. The code is presented as Supplementary Notes, but it has also been deposited in Dryad, along with all input files. Although the DOI provided in the text for the Dryad repository is not public yet, the reviewers can look at it using: <http://datadryad.org/share/LOcJ47Djnv98MkINgIDch-Qob-ujVTQt-edI-FLRiWE>

Reviewer #2 Remarks:

The report describes the development of a new version of the *Cryptosporidium* genome capture assay (CryptoCap_100k) for the diagnosis of *Cryptosporidium* species and mixed infections. Its performance in capturing *Cryptosporidium* genomes is extensively evaluated using in silico analysis and mock samples. Its ability to assign *Cryptosporidium* species and detect mixed infections is further evaluated using 100 clinical samples. The new assay is shown to have a broader detection range than the previous assay (CryptoCap_75k), has high sensitivity, and allows the detection of multiple *Cryptosporidium* in individual samples. It also allows pooling of samples during analysis, reducing the costs associated with genomic sequencing of *Cryptosporidium* spp.

Main comments:

The title of the report ("Whole Genome Targeted Enrichment and Sequencing of Human-Infecting *Cryptosporidium* spp.") and much of the introduction on research needs is about whole genome sequencing (WGS) of *Cryptosporidium* spp. in clinical samples. Therefore, the focus of the study is expected to be on the performance of CryptoCap_100k in WGS analysis of *Cryptosporidium* spp. This is especially the case when CryptoCap_100k is built upon the existing CryptoCap_75k and genome capture is a technology widely used in WGS analysis of other pathogens. Therefore, the report should focus on the utility of CryptoCap_100k for *Cryptosporidium* sequencing, with less emphasis on the development and in silico evaluation of CryptoCap_100k itself.

Thanks to the reviewer's feedback. We have done one minor edit and one major edit in this regard. To not sell the idea that CryptoCap_100k can in all cases recover the whole genome sequence of the sample, we have removed "whole" from our title. Second, to focus on the utility of CryptoCap_100k, we have summarized simulations and focused on analyses with clinical samples. We appreciate the reviewers help in us find a way to validate our approach!

The study could be much more vigorous with more data from experiments rather than in bioinformatics analysis and in silico simulations. Currently, two main tables (Tables 2 and 3) and two main figures (Fig. 1 and Fig. 2) from these analyses should be included in the supplementary materials. The associated

descriptions of these data have taken up a large part of the Results section (L130-L172) and should be shortened. In addition, Fig. 3, Fig. 4 and most of the supplementary materials are also based on *in silico* analyses and mock samples. Only Fig. 5 and Fig. 6 present data generated from 100 clinical samples. Since these two figures present similar data generated from two library preparation protocols, only one should be presented as the main figure. Table 4 is another table comparing the data from the 100 samples processed with two library preparation protocols.

Thanks to the reviewer's feedback. We followed the reviewer's suggestions to focus the main paper on the laboratory experiments and have moved most of the *in silico* portions to Supplementary Notes and Supplementary Results. As well as we have moved Figs. 1 and 2 to Supplementary Figures, and Table 3 to Supplementary Tables. We still believe it is valuable to retain Figure 3, now Figure 1, to provide comparisons when oocysts are used, rather than fecal samples. Additionally, we find value in comparing both library protocols (Figures 5 and 6, now Figures 3 and 4), as scientists can use both, and we highlight their differences.

In addition, the concluding statement at the end of the abstract needs to be supported by data generated with real samples rather than *in silico* simulations.

The last sentence of our previous abstract read:

CryptoCap_100k increases the percentage of reads mapping to target Cryptosporidium references in a various scenarios, increases the depth and breadth of genome coverage, and increases the reliability of detecting species within a sample, while decreasing costs, opening new opportunities to understand the biology of these important pathogens.

Although all these statements are true not only in *in silico* simulations but also in pure oocysts and, importantly, in clinical samples, for example, see lines 251–262. But we have clarified our conclusion in the abstract (see lines 41–45), and we hope the reviewers agree that our claims are drawn from the data.

There is a need for in-depth analysis of the data generated from the 100 clinical samples, especially the quality of the sequence data generated in relation to *Cryptosporidium* species and parasite load in the samples (as indicated by Ct values of qPCR analysis of the initial DNA rather than after library

construction), such as percent genome coverage, average fold coverage, and genome assembly statistics. Such data would allow us to understand the performance of the genome capture and sequencing approach. Currently, these data are not presented in the Results section. I see buried in the Supplementary Results that for the 100 clinical samples analyzed, the depth and breadth of genome coverage were 28.1X and 31.3%, respectively.

Thanks to the reviewer's feedback. We wanted to clarify that the C_T values presented in this study are indeed from the initial DNA and not after library construction. Thus, we have followed the reviewers' suggestions and have unburied these analyses from the Supplementary Figures and Supplementary Results, and have included them in the main manuscript, allowing for an association between parasite load in the samples and the quality of the mapping statistics obtained in our study.

This result is presented in a very confusing way in the Results (L291-293): the depth and breadth of genome coverage were on average 1.12X and 2.25%, respectively, per 100,000 reads (for better understanding, Fig. 5B can be replaced with Fig. S35, while Fig. 6B can be replaced with Fig. S40). The reason for this disappointing result should be fully discussed.

We understand the reviewer's perspective on this comment and how our presentation of the information led to confusion. We made the replacements that the reviewer recommends, in addition to replacing what is now figure 2 with non-normalized values. We agree with this replacement because, typically, most papers report the total depth and breadth of coverage, and further normalization of sequencing effort is not typically reported; thus, this allows for more direct comparisons with other datasets. We kept normalized values in the Supplementary Figures.

We briefly discuss what we attribute to obtaining an average breadth of genome coverage of 50% in lines 377–381.

The *Cryptosporidium* species identity of the initial 7 samples and the 100 clinical samples should be clearly stated, and samples with mixed species should be included. I could not find such information on the *Cryptosporidium*-positive samples anywhere in the main report and supplementary materials. Based on the data presented in Figure 5 and Figure 6, it appears that most of the 100 samples were from *C.*

parvum, with a few from *C. hominis*. As CryptoCap_100k also targets four other species such as *C. meleagridis* and *C. viatorum*, samples positive for these should also be included. Agreement and disagreement between genotyping and CryptoCap_100k results, especially regarding mixed infections, should be fully described and discussed.

Thanks to the reviewer for this comment. For the seven clinical samples, it is essential to note that the nature, for most of them, in terms of species and whether they are single or mixed infections, of these samples was unknown. Three of these were also included in the study by CryptoCap_75k (Khanan et al., 2024), allowing us to make comparisons in lines 401–408 of our Discussion.

For the 100 clinical samples, their provenance is from the UK, and our co-authors performed a multi-locus typing assay, the results of which are included in this study. We have now included this information in the Methods lines 630–634, and it is also presented in Tables S13–S15. We have included language throughout the manuscript to explicitly state the identity based on these previous results and how they compare to our study.

Regarding the inclusion of clinical samples that include other species, such as those not belonging to *C. parvum* and *C. hominis*, we did our best with the resources available. Unfortunately, we did not have access to clinical samples from these other species, given that over 90% of cryptosporidiosis cases in humans are caused by these two species (Xiao, 2010). We give evidence based on our results on why we expect the baits to perform well in other target species in lines 439–452 of the Discussion.

In particular, the data presented in Figures 5 and 6 suggest that many of the clinical samples had moderate proportions of *C. baileyi*, *C. tyzzeri*, and other species for which there is limited evidence that they infect humans (as stated in L307-308). Has this finding been validated using species-specific PCR? There is no detailed discussion of the validity of this observation.

The results presented in Figures 5 and 6, now Figures 3 and 4, suggest that for single-enrichments, clinical samples may have low proportions of *C. baileyi*, *C. tyzzeri*, and other species, but these have limited evidence of infecting humans. These patterns tend to dissipate with double enrichments. In this version, we explicitly state and validate that these results are compared with multi-locus typing assays performed in the UK, and our double-enrichments align with these. We suspect this may be due to low-

quality hits and/or contamination of the genome assemblies used as reference for these analyses, likely bacterial reads.

Since this report is about the development of CryptoCap_100k based on CryptoCap_75k, the comparison of the performance of the two in the detection of different species and mixed species should be done with real samples rather than in silico analysis. Again, it is difficult to accept that no samples of other targeted *Cryptosporidium* species were used in the evaluation of CryptoCap_100k.

This report aims to present CryptoCap_100k as a comprehensive bait set for the enrichment of human-infecting *Cryptosporidium*, while allowing enrichment pools with multiple samples, rather than individual libraries, for time- and cost-efficiency. We do not aim to demonstrate that CryptoCap_100k's performance is superior to CryptoCap_75K's. Additionally, our protocol is inspired by, but not based on, CryptoCap_75k, as our approach to bait design differs entirely. We have streamlined this language throughout the manuscript (e.g., lines 113–117).

Thanks to the reviewer's suggestion, we have also included in lines 401–408 a brief discussion of the similarities and differences obtained in three clinical samples for both methods.

Additional suggestions:

Some supplementary tables are nowhere to be found online, and all the Excel tables online are not labeled. In addition, it is very difficult to find the cited supplementary tables and figures. There are just too many of them. Some of them should be merged or deleted. Most journals have limits on the number of supplementary figures. Some of the supplementary tables in this report cannot be found online, probably because the number of supplementary items in this manuscript already exceeds the limit. These problems make it very difficult to review the manuscript.

We appreciate the reviewer highlighting these problems in the organization. To enhance readability, we have combined all supplementary tables into just one Excel file. We summarized the information from multiple tables that presented similar data in fewer tables (hence the naming changes), provided proper table titles and descriptions. We have removed all supplementary tables from the Supplementary Notes.

Similarly, we are providing all supplementary figures in a single stand-alone PDF file, accompanied by proper figure names and titles. We have removed all figures from the Supplementary Notes, where we now only note which chunks of code produce which figures.

Finally, we have deposited the R Markdown file that produced the Supplementary Notes on Dryad: <http://datadryad.org/share/LOcJ47Djnv98MkINglDch-Qob-ujVTQt-edI-FLRiWE>, along with all input files used in the study, to ensure the reproducibility of our code (the DOI for Dryad in the paper is not public until publishing). We hope that with these changes, organization and readability are enhanced, with supplementary figures, tables, and notes partitioned to contain different kinds of information.

Abstract and elsewhere (L77-81...), the challenge of sequencing *Cryptosporidium* genomes in clinical samples is overstated. In the era of single-cell sequencing, protocols for sequencing single oocysts purified from clinical samples have been published with good genome coverage. Without considering cost, most *Cryptosporidium* genomes in human fecal samples can be sequenced by purification of oocysts by immunomagnetic separation, enrichment of extracted DNA by whole-genome amplification, and second- and third-generation sequencing. Considering the cost of fabricating the genome capture device, the greatest advantage of the genome capture approach is the ability to sequence *Cryptosporidium* genomes in obtained DNA from epidemiologic and outbreak investigations.

We thank the reviewer for these comments, as they greatly helped us refine our theoretical approach. We agree with the reviewer's outlook, which summarizes that currently genome sequencing of *Cryptosporidium* involves, besides costs, additional methods such as oocyst purification. We have streamlined the introduction and discussion sections to better reflect the status of available technologies and their challenges. This includes the advantages and disadvantages of oocyst purification and whole-genome amplification, and thus why the development of other cost-effective alternatives for enriching target DNA can be introduced to the field to process large batches of samples. We hope the streamlining of this background throughout the paper now reads as accurate and adequate to the reviewer.

Introduction and L458-460. The problem of detecting mixed infections by PCR due to biased amplification also applies to CryptoCap_100k, as most capture baits are made from *C. parvum* sequences. If there is a mixed presence of *C. parvum* and *C. hominis* in a clinical sample, CryptoCap_100k would be expected to preferentially capture *C. parvum* sequences over *C. hominis* sequences. The failure to detect this in the computer simulation needs to be supported by vigorous testing with clinical specimens. Such statements should be toned down throughout the manuscript.

This is correct and is now highlighted in our results and discussion. For clinical samples, we incorporated analyses for mixed-infection samples, which proved to extend the value to CryptoCap_100k; however, we also alerted to potential biases related to *C. parvum*. Still, we present evidence of samples with full assignment to *C. hominis*, without any detectable bias towards *C. parvum*, in our population genomic analyses, indicating that these biases are relevant when *C. parvum* is in the mix. We hope these additional analyses provide evidence that, in addition to our simulations, we can still understand the composition of clinical samples while minimizing biases in clinical specimens.

The differences in the baits between CryptoCap_100k and CryptoCap_75k should be clearly stated. If the *C. parvum* baits in CryptoCap_100k were all redesigned without consideration of what was included in CryptoCap_75k (this could be true based on the description in the Methods), this should be clearly stated in early parts of the report. Such information is necessary to understand the differences between the two capture systems and their performance.

This is a great point; thank you for bringing it to our attention! Now we are more careful in stating that, compared to CryptoCap_75k, our study envisioned a different approach to bait design, without considering CryptoCap_75k (lines 111–117 of the introduction).

5. The methods section is difficult to follow. It would be helpful to have a diagram of the procedures used for library construction, genome capture, and Illumina sequencing, including whether capture and amplification were performed in plates and sequencing was performed on pooled products. Many readers are not familiar with the genome capture process used. Currently, readers have to figure it out by carefully

reading the entire Methods section. If whole genome capture was performed on metagenomic libraries generated from the extracted DNA instead of direct capture of sample DNA, the authors need to clearly state this early in the report.

Thanks for the suggestion! We have created Fig. 6, which presents the laboratory workflow used for clinical samples, to clarify our procedures. Additionally, we now explain, under *Hybridization Capture Enrichments with CryptoCap_100K*, that metagenomic libraries are pooled for enrichment.

6. The statement in L109-112 about including baits that capture subtelomeric regions of the genome in CryptoCap_100k is largely based on replacing IOWA-II with the better assembled IOWA-ATCC as the reference genome in the bait design (L378-379). While the IOWA-II genome is missing several telomeric ends, it has most of the subtelomeric sequences.

Thanks to the reviewer for pointing this out! Although the IOWA-II genome has some subtelomeric regions, the IOWA-ATCC assembly has the placement of all 16 subtelomeric regions and telomeres, which were considered in bait design (see Data Availability). We have also edited the language to highlight that our bait design uses genomes with **all** subtelomeric regions, rather than just subtelomeric regions. Hope these edits enhance the clarity behind our reasoning.

7. The length of the reads in statements in L285-286 should be provided. 150 or 250 PE?

Sorry, we missed providing this data for a few of our experiments in the previous version. Now we have taken a very careful look and provided this information for all experiments.

8. The writing is acceptable, but the abstract reads poorly with some language errors.

We have revised and rewritten the abstract, aiming to enhance readability and clarity. Thanks!

Below is the general note from the **Reviewer #1** in response to our **second** submission:

Reviewer #1 Remarks:

Thank you to the authors for their thorough and considered responses to my comments and suggestions. These edits and responses have addressed each of my queries, and I do not have further comments to add. My compliments to the authors on their high-quality work. I am happy to recommend publication.

Response to Reviewer 1:

Thank you for the compliments on the paper's scientific approach!

Below is the general note from the **Reviewer #2** in response to our **second** submission:

Reviewer #2 Remarks:

The authors addressed some of the issues I raised. Therefore, the revised report is an improvement over the previous submission. However, because no new experimental data are presented, my other suggestions remain unresolved.

Thank you for your feedback. While no new experimental data were generated, we have substantially revised the manuscript to clarify interpretations, reduce overstatements, and align results and conclusions. We hope we have done a good job addressing your concerns statement by statement.

Additionally, I still believe that some of the conclusions are too strong and not supported by the presented data. Results from *in silico* analyses, such as the data presented in Tables 1-3, should be moved to the supplementary materials.

We appreciated this perspective very much, as it really helped to ground our work, as we hope you will notice in our responses below.

Table 1 is not about *in silico* analyses; rather, it presents the species, genome sequences, and number of baits targeted for design to make CryptoCap_100K, and we believe it is important for the audience to understand our bait design approach, as it is one of the core approaches of the paper.

Regarding Table 2, we agree with the feedback. The information presented was about mapping analyses comparing the performance of CryptoCap_100K and CryptoCap_75K *in silico* (important but not essential to our paper), so it has been moved to Supplementary Tables.

Regarding Table 3, now Table 2, we think it is important to keep it in the main paper. This table is not about *in silico* analyses, but rather about the sequencing and mapping statistics from analyses of sequencing data from 100 patient samples prepared in the laboratory using two different library prep methods, highlighting differences with and without enrichment. We believe this table captures the value of our enriching assay for assessing human-infecting *Cryptosporidium* species from clinical/environmental samples, and, as authors, we feel it is valuable to our audience.

Below are some specific suggestions, listed in the order I read them rather than in order of importance:

- Abstract: The conclusions in the abstract should be modified. The data presented in this report do not support the conclusion that CryptoCap_100k "allows assessment of genetic variation via SNP calling." The relevant data presented in Figure 5 only support species identification. The nature of the clinical data presented, as well as the data presented in Figure 5, does not support identifying mixed species in the 91 clinical samples analyzed using the iNextEra library approach.

We agree that the SNP variants in Figure 5 (now Figure 4) are most useful for species identification. We demonstrate a broad array of additional analyses (e.g., estimates of phylogenetic distances, networks, fixation indices, etc.) from the SNPs called in Supplementary Figures (e.g., S61-S76 and S83-S88) and Tables (S17-S20). To more appropriately reflect this, we modified the statement in the abstract to: CryptoCap_100k “facilitates analyses of genetic variants in many samples...”.

- L37-40. It is general accepted that *C. tyzzeri* is not a human-infecting species, since there have been only a few reported cases of human infections. Please modify the statement and other related statements in the report.

Thanks to the reviewer for bringing this inaccuracy to our attention. We have modified our text in the places where we referred to *C. tyzzeri* as human-infecting to state “Here, we use updated genomic sequences from a broad diversity of human-infecting *Cryptosporidium* species that have been found to infect humans (*C. cuniculus*, *C. hominis*, *C. meleagridis*, *C. parvum*, *C. tyzzeri*, and *C. viatorum*).” We understand that *C. tyzzeri* is a rodent-infecting parasite, with only a few cases reported in humans. But given the latter, we hope this properly conveys that *C. tyzzeri* has been found to infect humans, although not primarily.

- L40-41, L56. Since spp. is a plural term, “*Cryptosporidium* spp. DNA” should be changed to “*Cryptosporidium* DNA”. Similarly, spp. should be followed by plural verbs.

Thanks! Changes were made that improved our text!

- L58. There are now far more than 39 recognized *Cryptosporidium* species.

Thanks! We made changes according to Ryan et al. (2021) and have included this new reference.

- L122. The word “data” is a plural term. Therefore, “This genomic data facilitates....” Should be replaced with “These genomic data facilitate....”

Thanks for bringing out this grammar error to our attention! Changes made.

- L133-136. This statement is difficult to understand. As it is written, the sentence implies that the comparison of the breadth of genome coverage between the two groups is made at different depths of coverage.

Thanks! We removed “at 1X” because, in the methods, we explained that breadth is always estimated at 1X depth.

- L134-135 and elsewhere. Keep one decimal place for percentage data.

Thanks for this comment, which prompted us to maintain consistency throughout the text.

Changes have been made to the main text, Table 2, and Supplementary Tables. With few exceptions, when percentage values were below 0.05 but larger than 0, we used two decimal places instead in the main text.

- L169-170. If the species assignment with 18S rRNA is only 70% correct, what is the rationale to include extra 18S rRNA baits in *CryptoCap_100k*?

Species identification in *Cryptosporidium* has typically been assessed by amplification and sequencing of the 18S rRNA gene. This practice is rooted in i) difficult access to whole genome data from clinical/environmental samples, ii) the presence of conserved regions amenable to universal primer design for this gene, iii) the presence of variable regions capable of distinguishing between species in this gene, and iv) in *Cryptosporidium* multiple copies of the gene are present, which enhances the sensitivity of the detection and aids in the resolution of closely related species (Turner et al., 2025) (this text has now been included in Supplemental Notes #4.3). Because 18S rRNA has been the marker of choice for this genus for decades, we wanted to incorporate baits that capture these regions for the future, as more data are gathered, to better understand the characteristics of this marker that underscore its uses and limitations and reflect the genetic complexity of these parasites. In summary, given that this is a standard marker in the system, we wanted to ensure we collected genetic data from these genes to support future comparisons of single gene vs. genomic trade-offs for species identification. That said, we are

hopeful that our approach, once applied to more diverse species, will support the identification of one or more single-copy markers that better serves species identification.

- L245-248. What are the implications for the observation of a significant proportion of reads from many clinical samples to non-parvum species, including those with no evidence of human infectivity? Are these results from the NEB-iTru libraries reliable? This appears not the case with data from iNextEra libraries

We agree with the reviewer, it is notable that, particularly in unenriched samples, many reads map to all genomes, with marked biases towards *C. baileyi* and *C. muris*. We suspect this may be due to i) likely contamination of the *C. baileyi* and *C. muris* genome sequences with likely bacterial reads, and ii) a high proportion of bacterial reads after library prep, particularly when libraries are not enriched with CryptoCap_100k, which map to these two genome assemblies (this text has now been included in Supplemental Notes #12.2). Now, when we focus on enriched libraries, we observe that reads map preferentially to the genomes to which they were assigned in the multi-locus typing system and our species assignment analyses, still with few but rare hits to *C. baileyi* and *C. muris* genome assemblies.

- L272-278. The presence of both *C. hominis* and *C. cuniculus* reads in a few samples in Fig. 4H could be due to the high genetic identity between these two species instead of mixed infections. This is correct, as observed in *in silico* simulations (Fig. S7), in *in vitro* simulations of single-infection samples (Fig. S17), and in the set of seven clinical samples (Fig. S32). Thanks to this comment, we realized we did not highlight this trend in the discussion, so we included a sentence in lines 392–397. Thanks!

- L289. Data from which library construction approach (see above) are used in the Population Genomics Analyses? Logically, data from double enriched iNextEra libraries should be used in such analyses. Figure 5 suggests that indeed this may be the case but descriptions in this section fail to reflect this.

This information is provided in the Methods, lines 697–699. The reviewer is correct, we used the data from the iNextEra libraries. Our Figure 3 legend also captures this library prep information.

- Discussion. This section should be significantly shorter. Currently, there is extensive repetition of results.

We appreciate the reviewer's suggestion to streamline the discussion. We reorganized the content to avoid repetition. We noticed that some intro paragraphs included elements also discussed elsewhere in the discussion. So, we made sure these elements were still discussed by moving sections from these paragraphs to other paragraphs, while removing elements that were repetitive. We also restructured the discussion, so we go from bait development to validation to application (subtitles used in the new discussion). The discussion was slightly shortened while retaining key interpretations necessary for the reader.

- Some statements about CryptoCap_100k's ability to perform SNP calling and comparative genomics should be modified. Similarly, the statements about detecting mixed species should be toned down.

Regarding SNP calling, we toned down some of our statements, which now, in the discussion, are written as: "CryptoCap_100k in our study was a useful tool for SNP calling, which enabled robust comparative analyses of genetic variation," and "The quality of the retrieved genomic data allows for SNPs to be called, utility for population genomics, and provides the potential to assess mixed-infections." These toned-down statements are supported by our methods, results, and discussion.

Regarding mixed infections, besides the toned-down statement highlighted above, we also toned down statements that now read, "...suggesting the potential for the genomic data obtained via CryptoCap_100k to support the deconvolution of mixed-infections within samples." And "However, these mixed-species and mixed-strain results should be interpreted cautiously and primarily as proof-of-concept, given the limited resolution of our dataset." Even our subtitle changed to "CryptoCap_100k Applicability: Potential for the Detection of Mixed Infections." We appreciate the reviewer's advice on helping us calibrate the reach of our study in our conclusions.

- Since data from the double-enriched iNextEra library approach appears more reliable, this approach should be the focus of the discussion and recommended. Currently, the authors suggest that both library approaches and single and double enrichments are suitable for species assignment, mixed species detection, and comparative genomic analysis.

Thanks to the reviewer's comment, now we are able to be more precise and suggest in the discussion "CryptoCap_100k substantially increased on-target reads relative to unenriched libraries (>437X fold change) for both NEB-iTru and iNextEra protocols. While NEB-iTru libraries still performed better than unenriched libraries, with a high percentage of reads mapping to target genomes, iNextEra libraries achieved greater genome coverage breadth at ~50%." to highlight there are key differences between the two. We also included in the conclusion paragraph "CryptoCap_100k can ... achieve as high as 3,690X fold-change enrichment with the iNextEra protocol in fecal samples" to ensure the audience understands that it is the more reliable approach. Given that there are no gold standards for library prep; thus, including this comparison among protocols is valuable and could serve as a benchmark for other studies on sequence capture enrichment assays.

- L357-364. These statements should be modified. Clearly, some species assignments from the NEB-iTru library approach are incorrect.

These lines refer to a paragraph that has now been removed due to excessive redundancy and overgeneralization. This comment was referring to the sentence we also commented on in the response above, which is now edited to read "CryptoCap_100k substantially increased on-target reads relative to unenriched libraries (>437X fold change) for both NEB-iTru and iNextEra protocols. While NEB-iTru libraries still performed better than unenriched libraries, with a high percentage of reads mapping to target genomes, iNextEra libraries achieved greater genome coverage breadth at ~50%." We hope this edit satisfies the reviewer, as it is more precise about which parameters both library prep methods behave similarly and which differ. Please note that in this statement, we don't talk about species assignment.

- As indicated above, the results shown in the study also do not support the conclusion that CryptoCap_100k sequencing enable SNP calling for population genomics.

We appreciate this concern and have revised the manuscript to clarify the scope of our conclusions. While our data support SNP calling and exploratory analyses of genetic variation—we were able to perform SNP calling and use it for population genomics in our set of 100 clinical samples—we are now being more careful in framing these results cautiously to avoid exaggerating their reach.

- Table 1 has no experimental data, therefore, should be moved to supplemental materials. Identity and overlap in the table should be defined. Without such information, it is difficult to understand the table.

Table 1 presents the number of baits designed from each reference genome sequence considered in CryptoCap_100k. We believe it is essential for the audience to understand our approach to bait design, as one of the paper's goals is to introduce this bait set to the community. Identity and overlap refer to the parameters used to cluster baits at various levels for bait design in VSEARCH, introduced in the Methods section, lines 487–489.

- Table 2. Genomes of *C. hominis* and *C. cuniculus* share greater than 98% sequence identity. It is unclear why the values of depth and breadth of genome coverages for *C. cuniculus* in the mapping of bait sequences are much smaller than those for *C. hominis*.

This is a great question! It is worth noting that these results refer to mapping both CryptoCap_100k and CryptoCap_75k baits to the twelve reference genome sequences, which show lower values for both approaches when mapped to *C. cuniculus*. We believe there could be different explanations for this behavior, but the one we gravitate towards the most is the curation of the reference genomes used here. *C. hominis* reference genomes are well-curated (RefSeq), with fewer and larger N50 contigs (1,445 contigs at 14.5 kb vs. 6,056 contigs at 1.8 kb) compared to the assembly from *C. cuniculus*, which is not curated. These differences translate into the reference being prone to fragmentation, so baits cannot map, missing subtelomeric regions (which we included baits from for *C. parvum*), and having more gaps and inconsistencies, which reflected poor mapping statistics.

- Tables 2 and 3. Table titles should have only a partial sentence. Other descriptions should be moved to footnotes.

Thanks for this suggestion! As recommended above, Table 2 was moved to the supplements. We have shortened the title to a partial sentence of Table 3 (now Table 2) and moved the descriptions to the footnotes.

- Figure 5. This figure has such low resolution that it is impossible to evaluate the accuracy of the associated descriptions in Results.

Thank you for raising this concern! We did not realize that Figure 5 showed poor resolution. We have further saved these and all main figures at a resolution of >300 DPI in the PDF and separately. These satisfy the journal requirements.

- Figure S1. The Y-axis should be on the same scale to allow for more accurate comparison between species.

Thank you for raising this point! When normalizing the y-axis, the signal for some genomes is lost because non-target species have many hits in the same regions. Thus, we normalized the y-axis to allow comparison and presented results with and without the target species to allow a closer look at the target species.