

Hidden diversity revealed: widespread *Trichuris incognita* reshapes understanding of human whipworm infections

Corresponding Author: Dr Pierre Schneeberger

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)

The current manuscript employs an amplicon sequencing approach to undertake population genetic study of *Trichuris* worms collected from humans from endemic countries in Africa and South East Asia. Through this analysis the authors reveal what may be a cryptic species that infects humans. The sample size is relatively small, but suggests this novel cryptic species, *T. incognita*, may make up a reasonable proportion of human infections (~15-20% prevalence in most of the study populations, with the exception of a higher prevalence in Uganda, albeit from small sample numbers). The authors explore their meta-data and do not identify any differences in the pathogenicity or infection demographics between the two populations.

Overall the work is well executed and the population genetics analyses are robust. However, the study is relatively small and in the absence of some major notable difference between the two parasite populations, I think this manuscript is suitable for a parasitology specific journal. At present the results are somewhat preliminary. Yes, the ITS2 data indicates two populations, but further studies, including morphological descriptions and, ideally, genomic studies, would be needed to confirm these are distinct species. There is definitely value in understanding the epidemiology of human helminthiases and in highlighting the importance of genetic studies of neglected parasites, noting that these may well be cryptic species complexes. However, the approaches deployed here are standard parasitological methods and the study, while interesting and well executed, does not provide insights that suggest a major and impactful finding.

Based on this, my opinion is that this manuscript is better suited to a specialist parasitology journal unless the authors can expand the study significantly and find that this novel species would have notable implications for global trichuriasis control.

Reviewer #2

(Remarks to the Author)

Dear editor,

The manuscript titled 'Nanopore-based analysis unravels the genetic landscape and phylogenetic placement of human-infecting *Trichuris* species in Côte d'Ivoire, Tanzania, Uganda, and Laos' investigates the presence of *T. trichiura* and *T. incognita* and infecting humans in the four countries and their genetics through the use of the nuclear ITS2 region. The manuscript presents interesting and comprehensive information regarding the lesser-known *T. incognita* and the *T. trichiura* in humans and is valuable in increasing our understanding of species in genus *Trichuris*. However, improvements in the data analysis, especially in the population structure, can be made. Comments for improvement are provided below:

1. Can the authors provide more information regarding why nanopore sequencing was selected as the technique of choice for this study? As only the ITS2 fragment was used, PCR coupled with sanger sequencing can be sufficient to investigate the identity and genetic diversity of *T. trichiura* and *T. incognita*.
2. Previous studies have also shown that length differences in the ITS2 sequences were useful for species differentiation between *Trichuris* species (e.g. *T. suis*, *T. ovis*, *T. leporis*). However, this is the first study demonstrating that length differences exist between the human-infecting *T. trichiura* and *T. incognita*, which is useful for human infections to distinguish between both species. Can the authors increase the emphasis regarding the importance of differentiating the

human-infecting *Trichuris* in the discussion section of the manuscript?

3. For population structure analysis in lines 131 onwards, analysis should be performed separately for *T. trichiura* and *T. incognita*. Only then we will be able to determine the differences between populations within *T. trichiura* and *T. incognita*.

4. Other population structure parameters that should be added include Tajima D's test, Fu and Li's test statistic, and Ka/Ks ratio. These parameters can help us to interpret the type of selection pressure, direction of selection, and evolution process occurring between and within populations.

5. The authors used percentage identity match (from BLAST) in their interpretation of results. Genetic distances may be a more appropriate measure for interspecies and intraspecies variation.

6. In line 373, the authors mentioned that a limitation of their study was the use of only the nuclear ITS2 marker for analysis. Mitochondrial genetic markers can also be added as an alternative marker of choice, aside from WGS.

Other minor comments:

1. Line 60. – *Trichuris* sp. should be changed to *Trichuris* sp.

2. Line 72 – The sentence can be changed to 'They have been shown to be a reliable marker to distinguish between *T. suis*, *T. incognita*, and *T. trichiura*.'

3. Line 77 – replace 'Low' to 'low'

4. Line 107 – take out (KK)

5. Line 113 – take out (ITS2)

6. Line 348 – *T.* should be changed to '*T. trichiura*'.

Reviewer #3

(Remarks to the Author)

I would like to thank the authors for their valuable contributions to research in neglected tropical diseases. Advancing our understanding of these diseases is crucial to improving control measures and working toward their elimination, particularly in the poorest and most affected communities around the globe. In this specific study, I appreciate the significant effort made by the authors and their collaborators to uncover the factors contributing to the low efficacy of anthelmintics in *Trichuris trichiura*. Additionally, their work to identify and disseminate findings on a new genetically defined group of *Trichuris* in Côte d'Ivoire is commendable. These findings, along with similar research, will provide critical insights into soil-transmitted helminth (STH) infections, their response to treatment, and their broader epidemiology.

However, I believe that substantial improvements are necessary before this work can be considered for publication in a scientific journal. My primary concern lies with the methodology, which is insufficiently described and/or justified. Below, I have outlined the key points of critique and suggestions for improvement to enhance the clarity and rigor of the study.

Major comments

The first point I would like to address concerns the bioinformatic pipeline used to obtain the ASVs. I have reservations about the use of *dada2* for ONT sequencing data. As I understand it, *dada2* is not ideally suited for processing nanopore sequencing information due to the platform's high error rate, and alternative tools have been specifically recommended (e.g., <https://doi.org/10.1186/s12864-024-10767-4>, <https://doi.org/10.1002/2211-5463.13868>). *Dada2* was originally designed for Illumina sequencing and later adapted for PacBio (<https://doi.org/10.1093/nar/gkz569>). Additionally, discussions on the *dada2* GitHub repository caution against its use with ONT data (e.g., <https://github.com/benjjneb/dada2/issues/1856>). I would like to inquire whether the authors have considered these limitations during their analysis and how they addressed the high error rates inherent to nanopore sequencing. I would recommend conducting a benchmarking assessment of other metabarcoding tools, such as *dadasnake* or *miniBarcoder*, and including this as supplementary material to substantiate the selection of the tool used for reliable ASV identification.

The recent paper by Venkatesan et al. (reference 16 of the manuscript) reported the presence of *T. incognita* only in Côte d'Ivoire, while the samples collected in Laos and Pemba were distinct. In contrast, this study identifies *T. incognita* across different countries. Highlighting and thoroughly discussing these seemingly contradictory findings is essential, particularly when both studies use the ITS2 marker, as the implications of either conclusion are significant. It is worth exploring whether these differences arise due to variations in sequencing techniques, bioinformatics pipelines, or the characteristics of the samples themselves. The study by Venkatesan analysed 87 stool samples, while this research examines a significantly larger set of 670 stool samples. Both studies cite the trial by Hürlimann et al. as the source of samples. However, upon examining the Patient ID or Sample ID coding, it appears there are discrepancies in how the samples are labelled or referenced. This raises a critical question: have both studies analysed the same set of samples? If so, it is crucial to assess whether the results derived from their respective methodologies align or diverge. Providing greater clarity on the comparability of the samples and methodologies used in these studies is essential. Discussing these points in depth would

enhance the reliability of the conclusions and shed light on the potential reasons behind the observed discrepancies.

The next point I would like to address is the filtering process of the sequences applied in the analysis. Given the importance of this study, which aims to identify a potentially new species on a global scale, I strongly believe that the authors should adopt a highly conservative approach to minimize the risk of reporting PCR or sequencing errors as *T. incognita*. This is especially critical if nanopore sequencing is the primary method used, due to its relatively higher error rates compared to other platforms. However, it seems the authors opted for a more relaxed set of filtering parameters, potentially increasing the likelihood of including erroneous sequences.

Specifically, I would like to comment on the use of the `maxEE` parameter in the `filterAndTrim()` function in `dada2`. Based on their GitHub code, the authors set `maxEE = 15`, with an annotation '`maxEE = 15 (15/500 = < 3% error rate)`'. However, it is unclear whether this parameter refers to an error threshold of 15 nucleotides per 500 bp, or 15 erroneous reads per 500 reads. Moreover, I am not sure that the criteria for selecting that threshold should be defined in that way, and in any case either interpretation seems excessively lenient, especially in the context of identifying a new species using ONT. To ensure greater confidence in the data, I recommend repeating the analysis with more conservative settings. For example, the default `maxEE = 2` is a reasonable starting point and is widely used in similar metabarcoding analyses. Additionally, using a stricter threshold, such as `maxEE = 1` (allowing no errors in filtered reads), could provide an even more robust basis for species identification. This stricter approach would align with the goal of minimizing the inclusion of erroneous reads, thereby enhancing the credibility of the findings. The results after using different thresholds for `maxEE` parameter can be added to the supplementary material, as I think can help to justify the methodology.

The work by Venkatesan employed additional stringent thresholds (in addition to a lower `maxEE`), including 'samples that had a total read depth of >1,000 mapped reads, >0.1% ASVs in the population overall, and >200 reads in an individual sample.' These thresholds, although arbitrary, reflect a conservative approach to minimize the inclusion of low-quality or potentially spurious data. In contrast, this study states that it obtained '198,792 reads (ranging from 5 to 9,420 high-quality reads per sample)' and 'ASVs below 5 reads across all samples were removed'. The lower threshold adopted in this work raises questions about its reliability. Filtering out ASVs with fewer than five reads may be too relaxed, as such low read counts are more prone to represent sequencing artifacts or background noise. A comparison between the two filtering strategies would benefit from a discussion of their implications for the final dataset and species identification. Furthermore, it is unclear whether the chosen thresholds in this work are supported by any previous studies or benchmarking analyses. Providing citations to validate these filtering criteria or explaining the rationale behind their selection is crucial for strengthening the study's methodological credibility.

Additionally, the lack of detailed metadata about the ASVs and their distribution across samples and countries detracts from the reproducibility and transparency of the findings. Supplementary materials, such as spreadsheets containing sample metadata and ASV information, could address this issue. I suggest including a spreadsheet as supplementary that includes the sample metadata relevant to this work and the information of ASVs similarly to the one included in this work: <https://doi.org/10.3389/frmbi.2023.1072059>; or here <https://doi.org/10.3389/fcimb.2024.1387126>. Thus, the information could be easily reanalysed.

The discrepancy between the number of ASVs and haplotypes in the manuscript raises several questions regarding the methodology and interpretation of the results. Conceptually, ASVs represent distinct sequences obtained through bioinformatics processing, while haplotypes are a set of closely linked genetic variations on a chromosome that tend to be inherited together. In ideal circumstances, the number of ASVs and haplotypes should be equal. However, the manuscript reports 474 ASVs and only 233 haplotypes, indicating potential methodological inconsistencies. The statement that 'alignment from good quality ITS2 ASVs was used for population genetic data and haplotype polymorphic analysis' suggests a quality control step that reduced the number of ASVs used in downstream analyses. However, the process by which the 'bad quality' of 241 ASVs was determined is unclear. Were these ASVs removed based on specific error rates, sequence length, or database matching criteria? Were those 'bad quality' ASVs indeed included in the phylogeny?

It is surprising that, except for Hap1, Hap2, Hap3, Hap4, Hap6, and Hap9, all the haplotypes displayed the same frequency. Does this frequency correspond to a single sequence or a fixed number of sequences? Interestingly, the haplotypes mentioned above (excluding Hap6) were the only ones with 100% identity in the databases, while the other 228 haplotypes were newly identified. This, combined with the uniform frequencies of over 200 haplotypes, raises the possibility that PCR and/or sequencing errors may have been inadvertently included in defining the haplotype network and diversity analyses. I recommend applying a more stringent sequence filtering process, as discussed earlier, and repeating the haplotype determination to ensure the robustness of these findings.

The nucleotide diversity (P_i) values reported in this manuscript are substantially higher than those observed by Venkatesan and colleagues using the same molecular marker, the ITS2, of comparable size. Specifically, Venkatesan reported P_i values of 0.000, 0.006, and 0.006 for Côte d'Ivoire, Laos, and Pemba, respectively. In contrast, this manuscript presents much higher P_i values of 0.236, 0.153, and 0.143 for the same regions. Additionally, the expected heterozygosity reported here is lower than that found in Venkatesan's study. Despite these significant differences, the manuscript does not address or discuss them, either in relation to Venkatesan's findings or within the broader context of previous work on helminth parasites, specifically using the ITS2 marker. Another notable discrepancy is the observation of higher P_i in Côte d'Ivoire in this manuscript, which contradicts the results of Venkatesan. Most studies associated with the clinical trial by Hürlimann have highlighted the higher selective pressure in Côte d'Ivoire due to its history of extensive albendazole and ivermectin coadministration for filarial control, which should have produced an impact on *Trichuris* spp. populations infecting humans. This selective pressure is generally expected to result in reduced genetic diversity; however, these findings show the

opposite trend, which warrants thorough discussion.

I would like to suggest a correction in the use of the random forest (RF) model for evaluating the predictive value of the fragment length in distinguishing species. First, I am uncertain if there was an issue with the file download, but I could not find any reference in the methodology to the use of the RF model or the association of the genetic data with the metadata. I request the authors clarify and detail the methodology employed for this analysis.

Although I am not an expert in machine learning (ML), my experience suggests that such approaches are typically implemented differently. In supervised ML, algorithms are trained with features to predict outcomes, building models capable of making accurate predictions about clinical or epidemiological variables. When no external validation data is available, the dataset is often divided into training (to build and optimize the model) and testing subsets (to evaluate its performance). Metrics like the Area Under the Curve (AUC) are calculated to represent predictive performance, and this process is repeated several times (e.g., through cross-validation) to obtain a mean and standard deviation for the AUC. Commonly, multiple methods, such as XGBoost, RF, or logistic regression, are compared, with the best-performing model (highest AUC) selected, optimized, and validated before reporting a final AUC. In this manuscript, the methodology for this analysis is absent, but from the text, it seems the RF model was applied to the entire dataset, reporting an AUC of 0.997. Such an extremely high AUC is rare in ML, even after optimization, and raises concerns that the predictive value may not have been assessed correctly. If the RF model was applied directly in the entire data, it suggests overfitting to the dataset, rendering the reported AUC invalid as a measure of predictive power that cannot provide additional information beyond a p-value. I recommend the authors collaborate with a data science expert to ensure proper application of ML methodologies if they intend to calculate the predictive value of the amplicon length accurately.

I would like to ask the authors about the metadata associated with the genetic data, which I find to be insufficiently described. The authors mentioned that their analysis focused on a subgroup of participants aged 6–18 years. As noted earlier, there is no detailed methodology explaining this analysis, so I would like to understand why the other participants were excluded from this analysis. Additionally, it would be valuable to see a description of the baseline characteristics of this subset to identify any potential biases introduced by the selection criteria. Furthermore, I suggest including an overall assessment of the differences in variables before stratifying by country. In these results, the authors reported differences; however, they did not explain how these differences were assessed, which statistical tests were used, or provide p-values. It seems that when one number was higher than another, it was simply stated without statistical validation. I strongly recommend that the authors clarify this section by specifying the statistical tests used, along with their respective alpha and beta values.

Finally, I would like to comment on the lack of efficacy data. All these samples were collected during a drug efficacy trial, yet the authors stated, 'the samples were collected within the context of clinical trials but these represented the status before drug use hence we did not explore how genetic diversity might relate to decreased drug efficacy'. I believe this omission represents a missed opportunity to link species-level differences to drug efficacy outcomes. For instance, were the samples identified as *T. incognita* associated with poorer treatment responses compared to those infected with *T. trichiura*? There are plausible reasons why the authors might not have conducted such analyses. Perhaps they lacked access to drug efficacy metadata or the permissions to disclose such data; however, they did report associations with other metadata, suggesting that this might not be the case. Another possibility is that they intend to use this data in a future publication. If so, I recommend that they state this explicitly in the discussion. In my view, linking drug efficacy data with genetic findings would have significantly strengthened the manuscript and facilitated acceptance. In any case, I request the authors explain why these analyses were omitted and clarify if such data are available. Presenting the frequency of *T. incognita* and *T. trichiura* in good versus poor responders to treatment would have provided a valuable and impactful addition to this work.

Minor comments

According to the script in the GitHub repository, some quality analyses are provided by DADA2 (e.g., via the functions `plotQualityProfile()` or `plotErrors()`). If the authors are committed to using DADA2, could they please include these quality assessment results in the supplementary material, especially if different filtering thresholds were evaluated?

The GitHub repository should include all the code used to prepare the manuscript. Currently, it appears to lack the scripts for generating the plots and the random forest analysis. Good bioinformatics practices require providing the complete code or scripts to ensure reproducibility, even for individuals with limited expertise. This work does not clearly adhere to those standards. I request that the authors update the GitHub repository with all relevant code.

The authors describe a study population of 670 participants (prior to PCR amplification), but only 629 samples yielded the amplicon. From my perspective, the true study population consists of the 629 participants whose stool samples successfully produced an amplicon and were included in downstream analyses. I suggest the authors redefine the study population accordingly.

Figure 3 is central to the paper, with much of the discussion revolving around its findings. However, I believe this figure could be improved for clarity. The external links within the figure could be made more intuitive. The "Host" and "Host type" categories appear redundant; I recommend merging them or retaining only "Host type" if more relevant. Additionally, consider combining the "Country" and "Continent" categories into a single "Region" class that specifies the country of origin within its respective continent. Also, you could use a colour code that helps to understand the figure (i.e. all European countries area coloured in blues). I also suggest adding a ring to indicate whether the sample originates from this study or another.

I find Figure 3b insufficiently explained and not informative. Could the authors provide a detailed explanation of what this figure depicts and its relevance to the results?

In my view, the population structure analyses are poorly described. The authors mentioned using DnaSP v6.0 and Arlequin v3.5.2 to calculate 'numbers of variable sites, number of haplotypes, nucleotide diversity (π), expected heterozygosity, and pairwise F_{ST} values.' However, it is unclear which software was used for each calculation or what parameter settings were applied. Please expand this section to include these details.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I thank the authors for their response to my previous review.

I don't wish to be obstructionist with the manuscript review and its really up to the editors to decide whether this study fits the journal brief. The review process asks me specifically what are the noteworthy results. The main findings of the study are that there are two distinction clades with human infectious *Trichuris* that have been designated as *T. trichiura* and *T. incognita*. However, as the authors point out in their rejoinder and in the manuscript (e.g., the discussion), *T. incognita* has already been described in the literature and differentiated from *T. trichiura* based on ITS sequence data. So this in itself is not a new finding.

The current study expands information on the prevalence of *T. incognita* in the specific study regions, which may support that it is a prevalent cause of human infections. This is interesting and definitely publishable data, but in the absence of data suggesting that this taxonomic difference has implications for human trichuriasis control (e.g, due to differences in epidemiology, reservoir hosts, drug tolerance or disease severity) it's difficult to assess whether this would be considered a highly impactful finding that warrants publication in a broad generalist journal. The authors expand on the potential value of these findings in the last few paragraphs of the discussion, but without any hard data indicating biological differences between *T. trichiura* and *T. incognita* that have public health relevance, this is very speculative. For example, the authors suggest potential relevance relating to the WHO's goal to eliminate trichuriasis as a public health problem by 2030. But unless they have data showing that *T. incognita* responds differently to treatment or has distinct transmission pathways to *T. trichiura* (e.g, a zoonotic reservoir) there is no data presently to suggest that the knowledge that *T. trichiura* may be a species complex doesn't really change the treatment and control effort.

As currently written, the paper seems to rely on the technical advance of deploying nanopore sequencing for ITS-2 typing. Using ITS-2 for nematode systematics is not a new approach (researchers have been doing this for decades). I don't see a major technical advance in deploying this alongside nanopore sequencing. It does not appear that the longer sequence reads are essential to be able to resolve the clades (which differ quite markedly over the 590bp amplicon) and the method isn't really amenable to field-deployment given it still requires a front-end PCR. So yes, they can use nanopore sequencing, and yes the approach appears scientifically sound, but I don't know that I can conclude that this alone is a major technical advance that one might consider high impact and noteworthy in the context of a publication in a generalist journal. Again, publishable? Yes. High-quality parasitology? Yes. But, in my view, suitable in a parasitology journal.

In my view, the interesting and noteworthy work that is consistent with a generalist journal is the, as far as I am aware, as yet unpublished data noted by the authors in their BioRxiv manuscript, describing the morphology and genomics of *T. incognita*.

Based on this, my opinion remains that the work is technically sound and the findings are scientifically justified, interesting and indeed publishable, but I do not see that, in the context of current knowledge in the field, this is a major advance that warrants publication in a generalist journal. However, ultimately that is a call for the journal editors and I wish the authors well with their research.

(Remarks on code availability)

Not applicable

Reviewer #4

(Remarks to the Author)

The use of nanopore-based long-read sequencing technologies represents an innovative and highly impactful approach, particularly in the field of nematode parasitology. Applying these techniques to study *Trichuris trichiura*, an important human parasite, the original research presented by authors is both timely and the amount of work done is commendable. The authors have effectively demonstrated how long-read methodologies can yield deeper and more accurate insights into the diversity, and evolutionary biology of parasitic organisms. Despite some initial concerns raised during the review process,

the authors have responded with obvious diligence. They carefully addressed all comments and suggestions from reviewers/editors, including those related to the sequencing methodology, such as sequencing depth and overall transparency.

Notably, they re-sequenced all original samples achieving a substantial increase in both sequencing depth (more than sufficient) and data quality. In addition, they implemented stricter quality control criteria and systematically resolved previous discrepancies throughout the manuscript. The deposition of standardized, publicly available code further enhances the reproducibility and transparency of their work.

In light of these major improvements and thoughtful responses to reviewer's prior concerns, I believe no further data analyses are necessary, the revised manuscript represents a significant contribution to the scientific community and should be considered for publication.

(Remarks on code availability)

The authors ensured that the analytical code is easily accessible and reusable by depositing it in a widely recognized public repository (GitHub). The project repository is well-structured, clearly described, and follows standardized code formatting, which significantly facilitates reproducibility and transparency. This level of accessibility is commendable and aligns with the current standards of open and reproducible science.

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/>

Point-by-point response:

We would like to thank the reviewers for the opportunity to revise and resubmit our manuscript. We believe the revisions have substantially improved the quality of the manuscript in response to the referees' valuable feedback. We therefore hope that it is now suitable for publication in *Nature Communications*.

Our responses are highlighted in blue.

Reviewer #1 (Remarks to the Author):

The current manuscript employs an amplicon sequencing approach to undertake population genetic study of *Trichuris* worms collected from humans from endemic countries in Africa and South East Asia. Through this analysis, the authors reveal what may be a cryptic species that infects humans. The sample size is relatively small, but suggests this novel cryptic species, *T. incognita*, may make up a reasonable proportion of human infections (~15-20% prevalence in most of the study populations, with the exception of a higher prevalence in Uganda, albeit from small sample numbers). The authors explore their meta-data and do not identify any differences in the pathogenicity or infection demographics between the two populations.

Overall the work is well executed and the population genetics analyses are robust. However, the study is relatively small and in the absence of some major notable difference between the two parasite populations, I think this manuscript is suitable for a parasitology specific journal. At present the results are somewhat preliminary. Yes, the ITS2 data indicates two populations, but further studies, including morphological descriptions and, ideally, genomic studies, would be needed to confirm these are distinct species. There is definitely value in understanding the epidemiology of human helminthiases and in highlighting the importance of genetic studies of neglected parasites, noting that these may well be cryptic species complexes. However, the approaches deployed here are standard parastilological methods and the study, while interesting and well executed, does not provide insights that suggest a major and impactful finding.

Based on this, my opinion is that this manuscript is better suited to a specialist parasitology journal unless the authors can expand the study significantly and find that this novel species would have notable implications for global trichuriasis control.

We thank the reviewer for their thoughtful feedback and for recognizing the robustness of our study and its contribution to understanding the epidemiology of human trichuriasis.

Regarding the study size, we acknowledge that our dataset represents a specific subset of *Trichuris* populations from selected endemic regions. Since the initial submission, we have expanded the number of samples processed at the PCR stage and, more critically, optimized our protocol to achieve substantially higher sequencing depth across samples. Specifically, we increased sequencing depth by more than 100-fold, which enabled the detection of previously unrecognized mixed infections, now shown to occur in over half of the samples in each country. These expanded findings significantly enhance the relevance of *T. incognita* detection and highlight its potential implications for both diagnostic accuracy and control strategies.

Our aim was to provide a broad-scale overview of *Trichuris* genetic diversity in human populations. We believe this represents a timely and valuable first step in detecting cryptic lineages and widespread co-infections, findings that are increasingly important considering reports of reduced treatment efficacy. While we agree that morphological and genomic studies would further strengthen species delimitation, we have conducted such analyses in a separate study focused on the discovery and characterization of *T. incognita* in Côte d'Ivoire, available here: <https://www.biorxiv.org/content/10.1101/2024.06.11.598441v4.full>. In this study, we further demonstrated that the two human-infecting *Trichuris* species are morphologically indistinguishable, yet exhibit pronounced genomic divergence.

This present study builds on those findings by exploring the geographical distribution of *T. incognita* and assessing whether its presence is localized or widespread—a distinction that has important implications for trichuriasis control.

Finally, while our molecular methods may be standard in parasitology, the integration of portable sequencing platforms such as the Oxford Nanopore platform represents a practical and innovative approach. This allowed us to resolve full-length ITS2 sequences, including clear species-specific length differences, which would likely be overlooked with short-read technologies. Furthermore, by combining DADA2-based variant detection, ANI-based clustering, and haplotype network analysis on Nanopore data, we demonstrate that accurate, high-resolution population genetics can be

achieved in a field-adapted, cost-effective manner. These advances lay the groundwork for future integrative studies and offer immediate relevance for diagnostics, surveillance, and control programs targeting human helminths.

Reviewer #2 (Remarks to the Author):

Dear editor,

The manuscript titled ‘Nanopore-based analysis unravels the genetic landscape and phylogenetic placement of human-infecting *Trichuris* species in Côte d'Ivoire, Tanzania, Uganda, and Laos’ investigates the presence of *T. trichiura* and *T. incognita* and infecting humans in the four countries and their genetics through the use of the nuclear ITS2 region. The manuscript presents interesting and comprehensive information regarding the lesser-known *T. incognita* and the *T. trichiura* in humans and is valuable in increasing our understanding of species in genus *Trichuris*. However, improvements in the data analysis, especially in the population structure, can be made. Comments for improvement are provided below:

1. Can the authors provide more information regarding why nanopore sequencing was selected as the technique of choice for this study? As only the ITS2 fragment was used, PCR coupled with sanger sequencing can be sufficient to investigate the identity and genetic diversity of *T. trichuria* and *T. incognita*.

We appreciate the reviewer’s insightful question regarding our choice of sequencing technology. We selected Oxford Nanopore sequencing (MinION/PromethION) due to its unique advantages for cost-effective, high-throughput sequencing, particularly in field-based and resource-limited settings where rapid and scalable genomic surveillance is essential.

Although Sanger sequencing remains the gold standard for analyzing single amplicons such as the ITS2 marker, it is inherently limited in throughput and scalability, particularly when handling large sample cohorts. Critically, it lacks the resolution needed to detect intra-sample genotype diversity—an essential aim of our study, given previous evidence of mixed *Trichuris* infections. In contrast, Oxford Nanopore sequencing enables parallel sequencing of numerous amplicons per sample and supports highly multiplexed workflows. This not only dramatically reduces per-sample

costs but also facilitates comprehensive population-level analyses, thereby providing the necessary resolution and scalability to address our research objectives.

Additionally, the portable nature of the MinION device and its ability to deliver real-time data acquisition offer major logistical advantages, including faster turnaround times compared to conventional sequencing platforms. Although our current study focuses on the ITS2 marker, the flexibility of Nanopore technology provides a future-proof platform that can readily accommodate extended genomic regions or multi-locus barcoding strategies in subsequent work.

2. Previous studies have also shown that length differences in the ITS2 sequences were useful for species differentiation between *Trichuris* species (e.g. *T. suis*, *T. ovis*, *T. leporis*). However, this is the first study demonstrating that length differences exist between the human-infecting *T. trichiura* and *T. incognita*, which is useful for human infections to distinguish between both species. Can the authors increase the emphasis regarding the importance of differentiating the human-infecting *Trichuris* in the discussion section of the manuscript?

We thank the reviewer for this excellent suggestion. We agree that the ability to distinguish between human-infecting *Trichuris trichiura* and *T. incognita* based on ITS2 sequence length is an important and novel finding of our study, with potential implications for diagnosis, epidemiological surveillance, and treatment monitoring. In response, we have revised the discussion to highlight the significance of this observation more clearly, particularly in the context of regions where both species may co-occur and contribute to treatment failure or misdiagnosis. For that, we have added the following text at lines 391-397:

“Placing these patterns into an evolutionary and public health context, our findings raise critical concerns. The frequent mixed infections, divergent population structures, and emergence of *T. incognita* in areas with suboptimal ivermectin-albendazole efficacy highlight the need for urgent reevaluation of species-specific transmission and treatment dynamics. With drug resistance already documented in veterinary whipworms^{30–32}, emerging evidence of external variables such as the gut microbiota influencing treatment efficacy³³, and signs of declining efficacy in human treatments¹⁸, the window to act preemptively is narrow.”

3. For population structure analysis in lines 131 onwards, analysis should be performed separately for *T. trichiura* and *T. incognita*. Only then we will be able to determine the differences between populations within *T. trichiura* and *T. incognita*.

We thank the reviewer for this important observation. We fully agree that population structure analyses should be performed separately for *T. trichiura* and *T. incognita* to accurately assess intraspecific variation. In response, we have re-analyzed the population genetic data for each species independently. These updated results are now presented in Table 2 and Fig. 4a–b, allowing for a clearer comparison of population structure within each species, without the confounding effects of inter-species divergence. For clarity and consistency, we grouped individuals by country within each species to perform these analyses.

4. Other population structure parameters that should be added include Tajima D's test, Fu and Li's test statistic, and Ka/Ks ratio. These parameters can help us to interpret the type of selection pressure, direction of selection, and evolution process occurring between and within populations.

We agree that additional population genetic statistics such as Tajima's D and Fu and Li's D and F can provide valuable insights into evolutionary dynamics. In response, we have now calculated these neutrality tests separately for *T. trichiura* and *T. incognita*, grouped by country, using DnaSP with default parameters based on the total number of mutations. The results are presented in the revised Results section and Table 2. These tests showed consistent patterns suggesting recent population expansion or limited long-term genetic divergence, particularly within *T. incognita*.

5. The authors used percentage identity match (from BLAST) in their interpretation of results. Genetic distances may be a more appropriate measure for interspecies and intraspecies variation.

We thank the reviewer for raising this important point regarding our use of BLAST percentage identity in the interpretation of results. In addition to the BLAST-based identity values (now provided in Supplementary Data 5), we conducted a genetic distance analysis to offer a more robust measure of both inter- and intraspecific variation. We have also updated the heatmap in Figure 5a to incorporate the enhanced sequencing data, which presents pairwise genetic similarities between the sequences

generated in this study and reference sequences from public databases. We have discussed the corresponding results in lines 267– 291.

6. In line 373, the authors mentioned that a limitation of their study was the use of only the nuclear ITS2 marker for analysis. Mitochondrial genetic markers can also be added as an alternative marker of choice, aside from WGS.

We thank the reviewer for this insightful suggestion. As noted in line 373, we acknowledge that the use of only the nuclear ITS2 marker represents a limitation of our study. While ITS2 is widely used in parasitology for species delineation due to its high interspecific variability and ease of amplification, we agree that incorporating mitochondrial markers such as *cox1* or *nad1* could provide enhanced resolution of genetic diversity and phylogenetic relationships. These markers offer advantages including maternal inheritance, absence of recombination, and typically higher mutation rates than nuclear loci. However, at the time of this study, other commonly used markers such as ITS1 and mitochondrial markers like *nad-1*, *nad-4* and *cox1* had notably lower amplification and recovery rates in our sample set, as we also reported in Venkatesan et al. (2025) in a subset of samples. This limited their utility for large-scale population studies and was a key reason for selecting ITS2 as the primary target for species differentiation and diversity analysis in our approach.

In future studies, we plan to integrate mitochondrial sequencing to complement our ITS2-based findings and provide a more comprehensive view of *Trichuris* evolutionary dynamics. Additionally, while whole-genome sequencing (WGS) offers the most detailed insight into species boundaries and population structure, it entails greater logistical and resource demands. We have now included this discussion in the revised manuscript's "Limitations" section (lines 415–424).

Other minor comments:

1. Line 60. – *Trichuris* sp. should be changed to *Trichuris* sp.

We addressed this comment in the revised manuscript.

2. Line 72 – The sentence can be changed to ‘They have been shown to be a reliable marker to distinguish between *T. suis* *T. incognita*, and *T. trichuria*.’

We addressed this comment in the revised manuscript.

3. Line 77 – replace ‘Low’ to ‘low’

We have corrected this typo in the manuscript.

4. Line 107 – take out (KK)

We addressed this comment in the revised manuscript.

5. Line 113 – take out (ITS2)

We addressed this comment in the revised manuscript.

6. Line 348 – T. should be change to ‘*T. trichiura*’.

We have corrected this typo in the manuscript.

Reviewer #3 (Remarks to the Author):

I would like to thank the authors for their valuable contributions to research in neglected tropical diseases. Advancing our understanding of these diseases is crucial to improving control measures and working toward their elimination, particularly in the poorest and most affected communities around the globe. In this specific study, I appreciate the significant effort made by the authors and their collaborators to uncover the factors contributing to the low efficacy of anthelmintics in *Trichuris trichiura*. Additionally, their work to identify and disseminate findings on a new genetically defined group of *Trichuris* in Côte d'Ivoire is commendable. These findings, along with similar research, will provide critical insights into soil-transmitted helminth (STH) infections, their response to treatment, and their broader epidemiology.

However, I believe that substantial improvements are necessary before this work can be considered for publication in a scientific journal. My primary concern lies with the methodology, which is insufficiently described and/or justified. Below, I have outlined the key points of critique and suggestions for improvement to enhance the clarity and rigor of the study.

We thank the reviewer for their thoughtful and constructive comments, which have greatly contributed to improving the quality and clarity of our manuscript. We fully acknowledge the concerns regarding the sequencing methodology, particularly the depth and transparency of our approach.

In response, we re-sequenced all samples from the initial study using four PromethION flow cells, resulting in a substantial increase in sequencing depth and data quality. We also implemented stricter data processing criteria throughout the analytical pipeline to enhance the robustness and reproducibility of our results. Specifically, we made the following key improvements:

- **Increased sequencing depth by more than 100-fold** and expanded the dataset to include 69 additional samples, bringing the total to $N = 687$. This also includes four positive single worm controls (*T. incognita* and *T. trichiura*). Most notably, this deeper sequencing markedly improved our ability to detect **mixed infections**, which were underrepresented in the previous version. While the main findings from the original submission were confirmed, the increased sequencing coverage revealed a substantially greater number of *T. incognita* infections, thereby strengthening the overall relevance and impact of our results.
- **Raised quality thresholds** across all key *in silico* processing steps to ensure more accurate and confident identification of amplicon sequence variants (ASVs).
- **Corrected and clarified methodological details** that were previously insufficiently described or ambiguous, as noted by the reviewer.

We believe these comprehensive revisions have substantially strengthened the methodological foundation of the study and significantly improved its scientific rigor, and would like to thank the reviewer once again for the in-depth assessment of our initial submission.

Major comments

The first point I would like to address concerns the bioinformatic pipeline used to obtain the ASVs. I have reservations about the use of dada2 for ONT sequencing data. As I understand it, dada2 is not ideally suited for processing nanopore sequencing information due to the platform's high error rate, and alternative tools have been specifically recommended (e.g., <https://doi.org/10.1186/s12864-024-10767-4>, <https://doi.org/10.1002/2211-5463.13868>). Dada2 was originally designed for Illumina sequencing and later adapted for PacBio (<https://doi.org/10.1093/nar/gkz569>).

Additionally, discussions on the dada2 GitHub repository caution against its use with ONT data (e.g., <https://github.com/benjjneb/dada2/issues/1856>). I would like to inquire whether the authors have considered these limitations during their analysis and how they addressed the high error rates inherent to nanopore sequencing. I would recommend conducting a benchmarking assessment of other metabarcoding tools, such as dadasnae or miniBarcode, and including this as supplementary material to substantiate the selection of the tool used for reliable ASV identification.

This is an important and technically thoughtful comment. We acknowledge the general limitations of applying DADA2 to Oxford Nanopore data, given the historically higher error rates associated with this platform. However, in our case, we achieved very high raw read quality using PromethION, with a median Q-score of 28+, corresponding to a per-base error rate of approximately 0.141% (i.e., 99.85% base-call accuracy), as now shown in **Supplementary Figure 2a**.

Given this high base accuracy, we opted to continue using DADA2 for amplicon sequence variant (ASV) inference, while incorporating substantially more stringent quality control and filtering steps throughout the pipeline. These steps are now detailed in our bioinformatics workflow (see **Supplementary Figure 1**) and elaborated upon in the revised Methods section. These include rigorous primer trimming, length filtering, and read quality thresholds tailored to ONT data characteristics. We also applied an additional “smoothing” step by clustering ASVs at a 98.5% identity threshold, based on the observed variability among positive controls (98.7%), in order to reduce potential inflation of diversity due to residual sequencing errors. Furthermore, we incorporated external reference sequences to filter out unrelated or spurious ASVs, ensuring that only taxonomically relevant sequences were retained in the final dataset.

To further ensure transparency and reproducibility, we have now included the complete ASV table and abundance matrix in **Supplementary Data 3**, which allows readers to assess the performance of the pipeline in detail.

We acknowledge the existence of alternative tools more specifically designed for ONT amplicon data (e.g., miniBarcode, dadasnae), and we appreciate the reviewer’s suggestion to benchmark these. While a full benchmarking comparison was beyond the scope of the current study, we are actively exploring these tools in ongoing work and plan to integrate such comparative assessments in future studies.

The recent paper by Venkatesan et al. (reference 16 of the manuscript) reported the presence of *T. incognita* only in Côte d'Ivoire, while the samples collected in Laos and Pemba were distinct. In contrast, this study identifies *T. incognita* across different countries. Highlighting and thoroughly discussing these seemingly contradictory findings is essential, particularly when both studies use the ITS2 marker, as the implications of either conclusion are significant. It is worth exploring whether these differences arise due to variations in sequencing techniques, bioinformatics pipelines, or the characteristics of the samples themselves. The study by Venkatesan analysed 87 stool samples, while this research examines a significantly larger set of 670 stool samples. Both studies cite the trial by Hürlimann et al. as the source of samples. However, upon examining the Patient ID or Sample ID coding, it appears there are discrepancies in how the samples are labelled or referenced. This raises a critical question: have both studies analysed the same set of samples? If so, it is crucial to assess whether the results derived from their respective methodologies align or diverge. Providing greater clarity on the comparability of the samples and methodologies used in these studies is essential. Discussing these points in depth would enhance the reliability of the conclusions and shed light on the potential reasons behind the observed discrepancies.

We thank the reviewer for this insightful comment. A detailed comparison with Venkatesan et al. (2024) is indeed warranted, given the shared use of the ITS2 marker and the partially overlapping sample sets. While Venkatesan et al. analyzed 87 stool samples and reported *T. incognita* only in Côte d'Ivoire (with distinct profiles in Laos and Pemba), our study includes a significantly larger dataset of 687 samples across four countries.

In addition, we employed long-read Nanopore sequencing, which enabled single-pass sequencing of full-length ITS2 amplicons (~530–600 bp), compared to the shorter Illumina paired-end reads (~460 bp combined) used in the previous study. The Illumina data had trimming applied at 240 bp for the forward read and 220 bp for the reverse read, potentially limiting resolution in variable regions of the ITS2 marker.

To ensure comparability, we harmonized sample ID coding and identified 40 overlapping samples between the two studies. Upon direct comparison, the results were largely consistent, particularly in the identification of the dominant *Trichuris*

species. However, our workflow also revealed additional cases of mixed infections, supported by increased sequencing depth and robust clustering fidelity.

Our updated pipeline incorporated the following key improvements:

- Increased sequencing coverage comparable to that in Venkatesan et al.,
- An upgraded and stringent *in silico* quality control pipeline,
- Positive single worm controls (*T. trichiura* and *T. incognita*) to calibrate clustering thresholds and identify sequencing artifacts,
- A 98.5% clustering identity threshold to reduce residual sequencing noise while preserving biological signal,
- A database-driven filtering approach to remove unrelated or non-target sequences.

Taken together, these methodological differences likely explain the broader geographic detection of *T. incognita* in our study. However, we emphasize that our findings are not contradictory to those of Venkatesan et al., but rather complementary, highlighting how sequencing platforms, bioinformatics pipelines, and analytical thresholds can influence the detection of genetic diversity. We now address this comparison in greater detail in the revised Discussion section of the manuscript.

The next point I would like to address is the filtering process of the sequences applied in the analysis. Given the importance of this study, which aims to identify a potentially new species on a global scale, I strongly believe that the authors should adopt a highly conservative approach to minimize the risk of reporting PCR or sequencing errors as *T. incognita*. This is especially critical if nanopore sequencing is the primary method used, due to its relatively higher error rates compared to other platforms. However, it seems the authors opted for a more relaxed set of filtering parameters, potentially increasing the likelihood of including erroneous sequences.

We thank the reviewer for raising this important point regarding sequence quality control and the risk of false positives, particularly in the context of species-level assignments using Nanopore data. We fully agree that a conservative and rigorous filtering strategy is essential to ensure the reliability of our findings, especially given

the taxonomic implications of identifying *T. incognita* across multiple geographic regions.

In the revised manuscript, we have implemented a highly conservative and multi-layered quality control pipeline, specifically designed to minimize the inclusion of PCR or sequencing artifacts. These steps are now outlined in detail in **Supplementary Figure 1** and summarized as follows:

1. Duplicate PCR reactions were performed for each sample to mitigate PCR amplification bias.
2. Basecalling was conducted using the latest Dorado model (v9.6 – v5 SUP).
3. Mean read quality filtering was applied, retaining only reads with Q scores ≥ 12 (the raw mean read quality was Q28+).
4. Length filtering was performed to retain reads between 520–700 bp.
5. Primer-based region extraction was conducted with zero mismatches allowed.
6. DADA2 denoising was applied with conservative parameters: minLen = 520, maxEE = 3, and maxN = 0.
7. Chimera removal was carried out using the removeBimeraDenovo function.
8. Abundance filtering excluded ASVs with a total relative abundance below 1% across all samples.
9. ASV smoothing by clustering: all raw ASVs were clustered using a 98.5% identity threshold, based on the observed 98.7% identity among single-worm positive controls, to remove residual sequencing noise while preserving true biological variation.
10. Taxonomic validation was performed by mapping the retained ASV centroids to a curated *Trichuris* reference database using Minimap2, with stringent alignment parameters to exclude off-target and contaminant sequences.

In addition to these steps, the median read quality score of the data entering the DADA2 pipeline was Q28+, corresponding to an estimated base-calling error rate of approximately 0.141%, comparable to Illumina amplicon sequencing. This, combined with strict abundance thresholds and the removal of rare variants and chimeras, ensures that only high-confidence ASVs are retained in downstream analyses.

We believe these measures provide a robust and conservative framework for accurate species-level identification and significantly reduce the risk of reporting PCR or sequencing errors as *T. incognita*.

Specifically, I would like to comment on the use of the maxEE parameter in the filterAndTrim() function in dada2. Based on their GitHub code, the authors set maxEE = 15, with an annotation 'maxEE = 15 (15/500 = < 3% error rate)'. However, it is unclear whether this parameter refers to an error threshold of 15 nucleotides per 500 bp, or 15 erroneous reads per 500 reads. Moreover, I am not sure that the criteria for selecting that threshold should be defined in that way, and in any case either interpretation seems excessively lenient, especially in the context of identifying a new species using ONT. To ensure greater confidence in the data, I recommend repeating the analysis with more conservative settings. For example, the default maxEE = 2 is a reasonable starting point and is widely used in similar metabarcoding analyses. Additionally, using a stricter threshold, such as maxEE = 1 (allowing no errors in filtered reads), could provide an even more robust basis for species identification. This stricter approach would align with the goal of minimizing the inclusion of erroneous reads, thereby enhancing the credibility of the findings. The results after using different thresholds for maxEE parameter can be added to the supplementary material, as I think can help to justify the methodology.

We thank the reviewer for this thoughtful and technically well-founded comment. In response, we have revised our pipeline to apply a more conservative maxEE threshold of 3, consistent with widely accepted practices in high-resolution amplicon sequencing. This change enhances confidence in retained reads and further reduces the risk of retaining erroneous sequences.

We acknowledge that our previously used setting of maxEE = 15 was overly permissive, originally based on the assumption that the low observed error rates relative to read length, in combination with DADA2's error model, would still allow accurate variant inference. However, we agree that this threshold is not appropriate for studies aiming at species-level resolution, particularly when using ONT data. With a median Q score ≥ 28 in our dataset, stricter quality filtering is both feasible and justified.

To further increase the robustness of species-level identification, we also implemented a 98.5% identity clustering step using Average Nucleotide Identity (ANI) via VSEARCH. This threshold was empirically derived from intra-species comparisons of ASVs generated from single-worm positive controls of *T. trichiura* and *T. incognita*, which were processed alongside study samples using the same stringent pipeline. The controls consistently showed $\geq 98.7\%$ pairwise identity among ASVs, supporting the use of 98.5% as a conservative and biologically meaningful cutoff. All ASVs passing quality control were subsequently clustered at this threshold, and representative centroids were used for downstream analyses. This additional step was critical to reduce the impact of residual sequencing noise while preserving true biological variation.

The work by Venkatesan employed additional stringent thresholds (in addition to a lower maxEE), including 'samples that had a total read depth of >1,000 mapped reads, >0.1% ASVs in the population overall, and >200 reads in an individual sample.' These thresholds, although arbitrary, reflect a conservative approach to minimize the inclusion of low-quality or potentially spurious data. In contrast, this study states that it obtained '198,792 reads (ranging from 5 to 9,420 high-quality reads per sample)' and 'ASVs below 5 reads across all samples were removed'. The lower threshold adopted in this work raises questions about its reliability. Filtering out ASVs with fewer than five reads may be too relaxed, as such low read counts are more prone to represent sequencing artifacts or background noise. A comparison between the two filtering strategies would benefit from a discussion of their implications for the final dataset and species identification. Furthermore, it is unclear whether the chosen thresholds in this work are supported by any previous studies or benchmarking analyses. Providing citations to validate these filtering criteria or explaining the rationale behind their selection is crucial for strengthening the study's methodological credibility.

In the initial version of the manuscript, our filtering thresholds were based in part on the substantial sequence divergence between the ITS2 regions of *T. trichiura* and *T. incognita* (~80% average nucleotide identity, as shown in **Figure 4a**). However, we fully agree that the earlier abundance and read depth thresholds were not sufficiently conservative, particularly for species-level assignments using Nanopore data, where sequencing errors and background noise can be more prevalent.

In the revised manuscript, we have significantly increased the stringency of our quality control pipeline, now applying thresholds that in several respects match or exceed those used in our earlier work (Venkatesan et al., 2024). Specifically, we have introduced a minimum read depth threshold of 500 reads per sample, a value commonly applied in metabarcoding and microbial community studies to ensure robust detection of biologically meaningful signal (e.g., [DOI: 10.1186/s40168-024-01993-9]). This threshold helps to eliminate low-yield samples that may otherwise contribute unreliable or spurious data.

Furthermore, we have retained only ASVs that pass a multi-step filtering and validation pipeline, including chimera removal, a 1% total abundance threshold, and 98.5% clustering based on empirically derived intra-species variation (as described in our previous response). This revised framework ensures a high-confidence ASV dataset that is both taxonomically reliable and biologically representative.

We believe these improvements not only address these concerns but also enable the detection of previously underappreciated patterns, such as the high frequency of co-infections, which are now more clearly resolved and discussed in the revised manuscript.

Additionally, the lack of detailed metadata about the ASVs and their distribution across samples and countries detracts from the reproducibility and transparency of the findings. Supplementary materials, such as spreadsheets containing sample metadata and ASV information, could address this issue. I suggest including a spreadsheet as supplementary that includes the sample metadata relevant to this work and the information of ASVs similarly to the one included in this work: <https://doi.org/10.3389/frmbi.2023.1072059>; or here <https://doi.org/10.3389/fcimb.2024.1387126>. Thus, the information could be easily reanalysed.

We thank the reviewer for highlighting the absence of this information. In response, we have now included detailed metadata for the 207 *Trichuris* ASVs identified in the study. This includes the representative nucleotide sequences (provided in **Supplementary Data 3**) as well as the associated sample metadata, including country of origin, sample ID, and relevant variables (**Supplementary Data 1**). These additions aim to enhance the transparency and reproducibility of our results.

The discrepancy between the number of ASVs and haplotypes in the manuscript raises several questions regarding the methodology and interpretation of the results. Conceptually, ASVs represent distinct sequences obtained through bioinformatics processing, while haplotypes are a set of closely linked genetic variations on a chromosome that tend to be inherited together. In ideal circumstances, the number of ASVs and haplotypes should be equal. However, the manuscript reports 474 ASVs and only 233 haplotypes, indicating potential methodological inconsistencies. The statement that ‘alignment from good quality ITS2 ASVs was used for population genetic data and haplotype polymorphic analysis’ suggests a quality control step that reduced the number of ASVs used in downstream analyses. However, the process by which the ‘bad quality’ of 241 ASVs was determined is unclear. Were these ASVs removed based on specific error rates, sequence length, or database matching criteria? Were those ‘bad quality’ ASVs indeed included in the phylogeny?

We thank the reviewer for raising this important point. We have now clarified the distinction between ASV clusters and haplotypes in the revised manuscript. Specifically, haplotypes generated via an additional level of clustering performed by DnaSP—were used solely for the haplotype network analysis. All other downstream analyses were conducted using the 207 high-quality ASV clusters.

To elaborate, in our updated dataset we initially identified over 2,800 ASVs following DADA2 processing. These underwent several layers of quality filtering and refinement to reduce sequencing noise and ensure biological relevance:

1. Minimap2 alignment against a curated *Trichuris* reference database was used to remove off-target and contaminant sequences, reducing the set to 2,275 high-quality ASVs.
2. To further improve accuracy and avoid overestimation of diversity, we applied consensus clustering at a 98.5% Average Nucleotide Identity (ANI) threshold using VSEARCH. This threshold was empirically derived from single-worm positive control specimens of *T. trichiura* and *T. incognita*, which were processed through the same stringent pipeline. These controls showed within-species identity of $\geq 98.7\%$, justifying the use of 98.5% as a conservative and biologically meaningful cutoff.

3. This clustering reduced the dataset to 215 consensus ASV clusters, from which 8 unclassified clusters were excluded due to lack of alignment to any known *Trichuris* reference.

The remaining 207 ASV clusters were used for all population-level and species-assignment analyses. For haplotype network reconstruction, these ASV sequences were aligned and processed using DnaSP, which detects polymorphic sites (e.g., SNPs, indels) and collapses identical sequences into haplotypes, where even a single nucleotide difference defines a new haplotype. This yielded 170 high-confidence haplotypes, each representing a distinct sequence group derived from the clustered ASVs.

We have clarified the stepwise reduction from raw ASVs to consensus ASVs to haplotypes in the revised Methods section, and we believe these updates enhance both the transparency and biological rigor of our analytical framework. Notably, the conversion from ASV clusters to haplotypes was applied specifically for the visualization of population structure and evolutionary relationships in PopART. For all other analyses, we retained the original ASV clusters to preserve the full granularity and resolution of the observed sequence variation.

It is surprising that, except for Hap1, Hap2, Hap3, Hap4, Hap6, and Hap9, all the haplotypes displayed the same frequency. Does this frequency correspond to a single sequence or a fixed number of sequences? Interestingly, the haplotypes mentioned above (excluding Hap6) were the only ones with 100% identity in the databases, while the other 228 haplotypes were newly identified. This, combined with the uniform frequencies of over 200 haplotypes, raises the possibility that PCR and/or sequencing errors may have been inadvertently included in defining the haplotype network and diversity analyses. I recommend applying a more stringent sequence filtering process, as discussed earlier, and repeating the haplotype determination to ensure the robustness of these findings.

We thank the reviewer for raising this important concern regarding the unusually uniform frequencies observed among most haplotypes in the earlier version of our analysis. As correctly noted, aside from Hap1, Hap2, Hap3, Hap4, Hap6, and Hap9, many haplotypes displayed identical frequencies, suggesting the potential inclusion of sequencing or PCR artifacts in haplotype inference.

We acknowledge that this pattern, along with the high number of ASVs initially reported, was derived from an earlier version of the dataset that employed a more permissive filtering and clustering strategy. In the revised manuscript, we have reanalyzed the data using a substantially more stringent pipeline, incorporating improved quality control and conservative thresholds, which resulted in a reduced and more biologically plausible total of 207 ASV clusters, and subsequently 170 haplotypes. This updated count aligns well with the observed population structure and genetic divergence in the dataset. Furthermore, we conducted a BLAST search of all 170 haplotypes, which revealed sequence identity ranging from 92% to 100% with previously described *Trichuris* species from humans, non-human primates, and unexpectedly pigs. These updated findings provide stronger support for the biological validity of the retained haplotypes and reduce the likelihood that they result from sequencing artifacts.

We have now fully integrated this reanalysis into the revised Results and Discussion sections, and we believe it significantly strengthens the robustness and interpretability of the diversity metrics presented in the manuscript.

The nucleotide diversity (P_i) values reported in this manuscript are substantially higher than those observed by Venkatesan and colleagues using the same molecular marker, the ITS2, of comparable size. Specifically, Venkatesan reported P_i values of 0.000, 0.006, and 0.006 for Côte d'Ivoire, Laos, and Pemba, respectively. In contrast, this manuscript presents much higher P_i values of 0.236, 0.153, and 0.143 for the same regions. Additionally, the expected heterozygosity reported here is lower than that found in Venkatesan's study. Despite these significant differences, the manuscript does not address or discuss them, either in relation to Venkatesan's findings or within the broader context of previous work on helminth parasites, specifically using the ITS2 marker. Another notable discrepancy is the observation of higher P_i in Côte d'Ivoire in this manuscript, which contradicts the results of Venkatesan. Most studies associated with the clinical trial by Hürlimann have highlighted the higher selective pressure in Côte d'Ivoire due to its history of extensive albendazole and ivermectin coadministration for filarial control, which should have produced an impact on *Trichuris* spp. populations infecting humans. This selective pressure is generally expected to result in reduced genetic diversity; however, these findings show the opposite trend, which warrants thorough discussion.

We agree that the initially reported nucleotide diversity (π) and heterozygosity values required revision and additional contextualization, particularly in light of the findings by Venkatesan et al. In response, we have re-sequenced the dataset using the Oxford Nanopore PromethION platform, enabling much deeper coverage and higher-resolution data across a substantially expanded sample set (687 individuals). Importantly, the updated analysis also incorporated the discovery of multiple ASVs across not only the two *Trichuris* species infecting humans (*T. trichiura* and *T. incognita*) but also a small number of sequences clustering with the swine-infecting clade. Given this expanded diversity, we conducted population genetic analyses separately for *T. trichiura* and *T. incognita* within and between the four countries to better resolve species-specific patterns.

The revised nucleotide diversity (π) estimates are as follows:

- For *T. incognita*: Côte d'Ivoire: 0.021; Laos: 0.012; Tanzania: 0.011
- For *T. trichiura*: Côte d'Ivoire: 0.017; Laos: 0.027; Tanzania: 0.012

While these values remain slightly higher than those reported by Venkatesan et al. (π = 0.000–0.006), we believe this difference is attributable to a combination of:

1. Increased sequencing depth via PromethION,
2. Inclusion of more ASVs per sample, following stringent quality control, and
3. Greater capture of intra-host diversity, including mixed-species and potentially mixed-strain infections.

We also acknowledge the reviewer's observation regarding selective pressure in Côte d'Ivoire from mass drug administration (MDA), which might be expected to reduce diversity. Interestingly, our updated dataset reveals that Côte d'Ivoire still harbors substantial genetic variation, with multiple co-circulating ASVs and evidence of mixed-species infections. This may reflect ongoing transmission, gene flow from untreated populations, or possibly recent introductions of divergent genotypes despite extensive MDA. We now discuss these alternative interpretations explicitly in the revised Results and Discussion sections, along with direct comparisons to Venkatesan et al.'s findings.

I would like to suggest a correction in the use of the random forest (RF) model for

evaluating the predictive value of the fragment length in distinguishing species. First, I am uncertain if there was an issue with the file download, but I could not find any reference in the methodology to the use of the RF model or the association of the genetic data with the metadata. I request the authors clarify and detail the methodology employed for this analysis.

Although I am not an expert in machine learning (ML), my experience suggests that such approaches are typically implemented differently. In supervised ML, algorithms are trained with features to predict outcomes, building models capable of making accurate predictions about clinical or epidemiological variables. When no external validation data is available, the dataset is often divided into training (to build and optimize the model) and testing subsets (to evaluate its performance). Metrics like the Area Under the Curve (AUC) are calculated to represent predictive performance, and this process is repeated several times (e.g., through cross-validation) to obtain a mean and standard deviation for the AUC. Commonly, multiple methods, such as XGBoost, RF, or logistic regression, are compared, with the best-performing model (highest AUC) selected, optimized, and validated before reporting a final AUC. In this manuscript, the methodology for this analysis is absent, but from the text, it seems the RF model was applied to the entire dataset, reporting an AUC of 0.997. Such an extremely high AUC is rare in ML, even after optimization, and raises concerns that the predictive value may not have been assessed correctly. If the RF model was applied directly in the entire data, it suggests overfitting to the dataset, rendering the reported AUC invalid as a measure of predictive power that cannot provide additional information beyond a p-value. I recommend the authors collaborate with a data science expert to ensure proper application of ML methodologies if they intend to calculate the predictive value of the amplicon length accurately.

We thank the reviewer for this important comment and fully agree that machine learning methods should be implemented with appropriate validation to avoid overfitting and ensure robust conclusions. In response, we have re-run the Random Forest (RF) model using the updated dataset, this time implemented leave-one-out cross-validation (LOOCV) via the caret package in R (Kuhn, 2008). This method provides a reliable estimate of model performance in settings with limited sample size

and helps prevent overfitting by ensuring that each data point is used exactly once for validation.

The model was trained to classify *T. trichiura* vs. *T. incognita* using amplicon fragment length as the sole predictor. As expected, fragment length emerged as a particularly strong discriminative feature, with *T. trichiura* averaging ~530 bp and *T. incognita* averaging ~600 bp. As expected, since there was no overlap between the fragment length from *T. trichiura* and *T. incognita*, the performance of the model was “perfect”. While we acknowledge that using machine learning may be somewhat excessive in this case—and that a simple length threshold could suffice—we chose to retain it for its systematic approach and to ensure that sequences derived from future studies, which may exhibit greater length variability, can still be accommodated by the prediction model. We also provided a confusion matrix to show that there was no class disbalance in the model. To ensure full transparency and reproducibility, we have made the complete R code and analysis pipeline publicly available via our GitHub repository: https://github.com/STPHxBioinformatics/HITS2/blob/main/ITS2_Trichuriasis_post_processing-main/ITS2_Trichuriasis_post_processing-main/Random_forest

These updates are now described in the revised Methods and Results sections.

I would like to ask the authors about the metadata associated with the genetic data, which I find to be insufficiently described. The authors mentioned that their analysis focused on a subgroup of participants aged 6–18 years. As noted earlier, there is no detailed methodology explaining this analysis, so I would like to understand why the other participants were excluded from this analysis. Additionally, it would be valuable to see a description of the baseline characteristics of this subset to identify any potential biases introduced by the selection criteria. Furthermore, I suggest including an overall assessment of the differences in variables before stratifying by country. In these results, the authors reported differences; however, they did not explain how these differences were assessed, which statistical tests were used, or provide p-values. It seems that when one number was higher than another, it was simply stated without statistical validation. I strongly recommend that the authors clarify this section by specifying the statistical tests used, along with their respective alpha and beta values.

We thank the reviewer for this insightful comment. Regarding the metadata and subgroup selection, we would like to clarify that samples from Uganda were collected exclusively from school-aged children, whereas the other countries included a broader age range. To ensure comparability across study sites and minimize potential age-related confounding, we performed key analyses using an age-normalized subset (ages 6–18) across all countries. This age group was selected because it reflects a biologically and epidemiologically relevant window in which *Trichuris* infections are most common and treatment efficacy is often assessed. This approach also allowed us to better examine physiologically meaningful metrics, such as infection intensity, sex-based differences, and population-level diversity patterns. To increase transparency, we have now included detailed baseline characteristics of the age-subsetted populations in **Supplementary Data 6**, including demographic distributions by country, age, and sex. Additionally, we have specified the statistical tests used for comparisons (e.g., Mann-Whitney, Kruskal-Wallis, or Fisher's exact test as appropriate) in the methods section with *p*-values reported where relevant. We believe these revisions enhance the rigor and interpretability of our demographic and genetic analyses and help contextualize the findings within participant-level metadata.

Finally, I would like to comment on the lack of efficacy data. All these samples were collected during a drug efficacy trial, yet the authors stated, 'the samples were collected within the context of clinical trials but these represented the status before drug use hence we did not explore how genetic diversity might relate to decreased drug efficacy'. I believe this omission represents a missed opportunity to link species-level differences to drug efficacy outcomes. For instance, were the samples identified as *T. incognita* associated with poorer treatment responses compared to those infected with *T. trichiura*? There are plausible reasons why the authors might not have conducted such analyses. Perhaps they lacked access to drug efficacy metadata or the permissions to disclose such data; however, they did report associations with other metadata, suggesting that this might not be the case. Another possibility is that they intend to use this data in a future publication. If so, I recommend that they state this explicitly in the discussion. In my view, linking drug efficacy data with genetic findings would have significantly strengthened the manuscript and facilitated acceptance. In any case, I request the authors explain why these analyses were omitted and clarify if such data are available. Presenting the frequency of *T. incognita* and *T. trichiura* in

good versus poor responders to treatment would have provided a valuable and impactful addition to this work.

We sincerely thank the reviewer for this insightful comment regarding the potential to link species-level *Trichuris* differentiation with drug efficacy outcomes. We fully agree that such analyses could provide valuable insights into treatment response variability and would indeed strengthen the translational impact of the study. However, as analyses of the primary clinical trial are still ongoing, including data consolidation and curation, we did not perform additional analyses to investigate associations between species identity and treatment efficacy.

To address the reviewer's concern and ensure transparency, we have revised the Discussion section to explicitly acknowledge this limitation and to clarify that treatment outcome data were not analyzed in the present study (see lines 421–423 of the revised manuscript).

Minor comments

According to the script in the GitHub repository, some quality analyses are provided by DADA2 (e.g., via the functions `plotQualityProfile()` or `plotErrors()`). If the authors are committed to using DADA2, could they please include these quality assessment results in the supplementary material, especially if different filtering thresholds were evaluated?

We thank the reviewer for this helpful suggestion and fully agree that visualizing quality assessments generated by DADA2 enhances the transparency and reproducibility of the sequencing workflow. In the revised supplementary materials, we have now included key DADA2 diagnostic plots, including `plotErrors()`, for representative samples across sequencing runs. These plots illustrate the accuracy of the error model learned during the denoising step. The selected plot has been added as **Supplementary Figure 3** in the revised manuscript.

The GitHub repository should include all the code used to prepare the manuscript. Currently, it appears to lack the scripts for generating the plots and the random forest analysis. Good bioinformatics practices require providing the complete code or scripts to ensure reproducibility, even for individuals with limited expertise. This work does not

clearly adhere to those standards. I request that the authors update the GitHub repository with all relevant code.

We thank the reviewer for this important observation. We fully agree that reproducibility is a cornerstone of robust bioinformatics research and have taken immediate steps to address this concern. The GitHub repository has now been updated to include all relevant scripts used in the preparation of the manuscript, including those for figure generation and for conducting the random forest classification analysis. This update ensures that all analyses presented in the study can be replicated and independently verified.

The authors describe a study population of 670 participants (prior to PCR amplification), but only 629 samples yielded the amplicon. From my perspective, the true study population consists of the 629 participants whose stool samples successfully produced an amplicon and were included in downstream analyses. I suggest the authors redefine the study population accordingly.

We appreciate the reviewer's thoughtful comment and agree that the study population should be defined based on the number of samples that successfully produced an amplicon and passed all downstream quality control steps. During the revision process, we reamplified samples that initially showed no visible bands or had insufficient read depth, allowing us to recover some previously excluded samples. We also included additional samples at this stage, bringing the total number of processed stool-derived samples to 687.

In addition, positive control worm samples were sequenced alongside the main dataset to calibrate clustering thresholds and validate species-specific profiles. We have updated the manuscript to reflect that the final dataset includes 687 stool samples and have applied this definition consistently throughout the revised text.

Figure 3 is central to the paper, with much of the discussion revolving around its findings. However, I believe this figure could be improved for clarity. The external links within the figure could be made more intuitive. The "Host" and "Host type" categories appear redundant; I recommend merging them or retaining only "Host type" if more relevant. Additionally, consider combining the "Country" and "Continent" categories into a single "Region" class that specifies the country of origin within its respective

continent. Also, you could use a colour code that helps to understand the figure (i.e. all European countries area coloured in blues). I also suggest adding a ring to indicate whether the sample originates from this study or another.

Thank you for your thoughtful and constructive feedback regarding **Figure 3**. We agree that this figure plays a central role in the manuscript, and we have implemented several revisions to enhance its clarity and interpretability based on your suggestions.

First, we streamlined the visual structure by merging the “Host” and “Host type” categories into a single “Host type” ring, eliminating redundancy and simplifying interpretation. We also consolidated the previous “Country” and “Continent” rings into a unified “Region” ring, which now presents both the country of origin and its continental grouping using hierarchical labelling and a continent-based color scheme. This improves visual coherence and enhances the geographic context of the samples.

Additionally, we added an outer ring to indicate whether each sample originated from the present study or from previously published datasets. This provides important context for understanding diversity patterns and species distributions.

Finally, we revised the figure layout, external link structure, and legend annotations to improve overall readability and accessibility for the reader.

I find Figure 3b insufficiently explained and not informative. Could the authors provide a detailed explanation of what this figure depicts and its relevance to the results?

We appreciate the opportunity to clarify the purpose and interpretation of **Figure 4 (Previously Figure 3b)**. This analysis is intended to provide an unsupervised visualization of sequence similarity, independent of phylogenetic assumptions, and to highlight genetic clustering patterns that may reflect geographic or taxonomic structure.

Specifically, Figure 4 complements the phylogenetic tree by presenting a matrix-based view that captures subtle sequence-level differences across the dataset. It is particularly useful for identifying clusters and examining their pairwise nucleotide identity relationships.

We agree that the original version of the figure lacked sufficient explanation. In the revised manuscript, we have added a dedicated section in the Results to describe

Figure 4 in greater detail and to clarify its role in interpreting genetic diversity and structure within and between *Trichuris* populations. Additionally, we have revised the figure legend to provide more explicit guidance for interpreting the heatmap and associated clustering dendrogram. We hope these revisions enhance the clarity and usefulness of the figure for readers.

In my view, the population structure analyses are poorly described. The authors mentioned using DnaSP v6.0 and Arlequin v3.5.2 to calculate 'numbers of variable sites, number of haplotypes, nucleotide diversity (π), expected heterozygosity, and pairwise F_{ST} values.' However, it is unclear which software was used for each calculation or what parameter settings were applied. Please expand this section to include these details.

Thank you very much for this helpful comment. We agree that the description of the population structure analyses required further clarification, and we have expanded this section in the revised manuscript to specify the tools and parameter settings used for each metric.

Specifically, the number of haplotypes, nucleotide diversity (π), number of segregating sites, and other haplotype polymorphism statistics were computed using DnaSP v6.0 with default settings. Pairwise F_{ST} values and their associated significance tests were calculated using Arlequin v3.5.2, with 1,0000 permutations applied to assess statistical significance.

These clarifications have now been incorporated into the Methods section of the revised manuscript.