This document is the reply from the authors to the reviewer's comments, it lists all modifications that were applied to account for the review. Please note that:

- Text in brown is a plain copy of the reviewer comments, text in dark green is the answer issued by the authors
- The line numbers for "initial text" refer to the first submitted version of the manuscript, those for "new text" refer to the revised version
- The new manuscript contains additional minor changes that are not listed here, but that we found useful to apply while re-reading it

A/ Review by Ali Hakimzadeh, 02 Apr 2024 12:33

Abstract

1/ The abstract provides a clear and concise overview of the mbctools package, its functionalities, and its advantages. However, there are several areas where clarity and detail could be enhanced:
Line 21-27: It is commendable that the tool is designed for users without command-line expertise, but it would be beneficial to briefly mention the specific user interface design or examples of how the menu-driven program simplifies the process.

Author response: As suggested by the reviewer, we have clarified the text explaining how the menu driven interface simplifies the process.
Text added to the revised manuscript (L33-37): "The software, designed to run in a console, offers an interactive experience, guided by keyboard inputs, assisting users along the way through data processing and hiding the complexity of command lines by letting them concentrate on selecting parameters to apply in each step of the process."

2/ Line 28-32: While VSEARCH's utilization is noted, a brief explanation of why VSEARCH was chosen over other tools, considering its advantages in processing amplicon data, would provide a more comprehensive background.

Author response: As suggested by the reviewer, we have clarified the choice for VSEARCH.
Initial text (L28): "VSEARCH is utilized for processing fastq files derived from amplicon data."

Revised text (L37-43): “In our workflow, VSEARCH is utilized for processing fastq files derived from amplicon-based Next-Generation Sequencing data. This software is a versatile open-source tool for processing amplicon sequences, offering advantages such as high speed, efficient memory usage, and the ability to handle large datasets. It provides functions for various tasks such as dereplication, clustering, chimera detection, and taxonomic assignment. VSEARCH is thus very efficient in retrieving the overall diversity of a sample.”

Introduction

3/ The introduction effectively sets the stage by highlighting the importance and applications of eDNA metabarcoding. It also identifies a gap in the availability of user-friendly bioinformatics tools. Nevertheless, some aspects could be refined for improved clarity and engagement:
Lines 39-52: The introduction to eDNA metabarcoding and its applications is well-articulated. However, referencing specific studies (1-3, 4, 5, 6-8, 9, 10) without any context may leave readers unfamiliar with the cited works without a clear understanding. Brief descriptions of these references could enhance the introduction's informativeness.

Author response: We agree with the reviewer’s assessment, we have added more context to the citations to improve the comprehensiveness of the introduction, we have added an entire paragraph to develop the topics for the 1-3, 4, 5, 6-8 citation and for 6 to 8 we have added some context.
Initial text (L39-52): “Over the past decade, environmental DNA (eDNA) metabarcoding has emerged as a robust and increasingly popular approach. Its simplicity of implementation enables biologists, local governments, and NGOs to increase their understanding of the spatial and temporal ecological networks (1–3). As illustrated in Figure 1, this method consists in identifying a targeted subset of genomes within bulk samples by massive sequencing of amplicons of taxonomically informative genetic markers, referred to as loci in mbctools (4,5). This approach has proven to be useful for diet profile analysis, air and water quality monitoring, biodiversity monitoring, and food quality control, beverage or ancient ecosystem composition (6–8). In our recent studies, we have employed this approach to investigate the transmission cycles of vector-borne diseases, using the midgut contents of vector insects as bulk samples, and simultaneously sequencing markers allowing for the identification of their genetic diversity, their diet composition (blood meal sources), their midgut microbiome composition, and pathogen diversity, as all these components and their interactions shape transmission cycles (9,10).”
Revised text (L52-94): “Metabarcoding consists in identifying a targeted subset of genomes within bulk samples by massive sequencing of amplicons of taxonomically informative genetic markers, also known as barcodes (Figure 1) (1). Over the past decade, this approach has quickly gained popularity since Hebert et al. (2) first advocated the use of short variable DNA sequences, amplified using universal primers, for species identification, and discovery of new taxa. Its simplicity of implementation enables biologists, local governments, and NGOs to increase their understanding of spatial and temporal ecological networks (3). Using universal

primers specific of kingdoms or subgroups of organisms of interest (e.g. "arthropods", "vertebrates", "reptiles", "amphibians", "mammals", "birds", ...), metabarcoding is highly effective for species-level identification and assessment of the whole diversity of a sample within the targeted kingdoms or subgroups of organisms of interest. For example, universal primers for vertebrates, targeting the 12S rRNA gene, have been proposed for metabarcoding studies focused on vertebrates (4). In studies targeting animals, primers allowing the amplification of a portion of the mitochondrial marker Cytochrome oxidase 1 (COI) may be used (5). In plants, standard DNA barcoding generally involves one to four plastid DNA regions (rbcL, matK, trnH-psbA, trnL), sometimes in combination with the internal transcribed spacers of nuclear ribosomal DNA (nrDNA, ITS) (6). For fungi, the internal transcribed spacer (ITS) regions ITS1 and ITS2 of the nuclear ribosomal DNA are the most commonly used barcodes (7). For bacterial and archaeal communities, the 16S ribosomal RNA (16S rRNA) is the most widely used barcode and provides taxonomic resolution to the genus level (8). Beyond these markers, specific markers are used in various fields to genotype species. For instance, in the study of Trypanosoma cruzi, the causative agent of Chagas disease, the Glucose-6-Phosphate Isomerase (GPI) and the Cytochrome oxidase 2 (COII) genes are frequently used to distinguish between different strains (9–11). Although single markers are very informative, in many cases, a combination of markers is necessary to fully understand the spatial and temporal dynamics of ecological networks. A huge advantage of metabarcoding is that it is based on massive sequencing. As a consequence, all the amplicons amplified from different targeted markers can be pooled and sequenced simultaneously (12). This powerful approach has proven to be useful for diet profile analysis, air and water quality monitoring, biodiversity monitoring, and food quality control, e.g., beverage or ancient ecosystem composition (13–15). As another application example, our group has recently proposed to use this approach to untangle the transmission cycles of vector-borne pathogens and their dynamics, using the gut contents of vector insects as bulk samples, and simultaneously sequencing markers allowing for the molecular identification of vector species and/or genetic diversity, blood meal sources (using universal primers for vertebrates which may also serve as pathogen hosts), gut microbiome composition (which may modulate vectorial capacity), and pathogen diversity. Indeed, all the latter components interact together to shape transmission cycles (12,16).”

4/ Lines 53-64: This section does well in identifying a gap in current bioinformatics tools. Briefly addressing the specific difficulties that novice users have using current tools would strengthen it and make the case for the development of mbctools more compelling.

Author response: We added the suggested content, and developed on the specific difficulties strengthening the need for a tool like mbctools.
Initial text (L39-52): “In contrast, mbctools has been designed with a simplified interface, allowing scientists without advanced technical expertise in bioinformatics to easily install, navigate and utilize the pipeline.”
Revised text (L103-108): “Indeed, the obtaining of pseudo quantitative taxonomic data from *fastq* files involves a series of complex processing steps that can be achieved with specific software implying the understanding of a large panel of parameters. To address this problem, *mbctools* was designed with a simplified interface, allowing scientists without advanced

scripting skills to easily install, navigate, utilize the pipeline while focusing on key parameters."

5/ Lines 65-78: The rationale for choosing VSEARCH and the description of mbctools' functionality are clear. However, stating that "the software can be used only with command lines" (Line 77) seems to contradict the abstract's emphasis on a user-friendly, menu-driven interface. Clarification on this point would be helpful. Additionally, illustrating how mbctools compares to or improves upon existing tools in sensitivity, specificity, or user-friendliness would add valuable context for the reader.

Author response: Thank you for pointing out some ambiguities in our manuscript. We have clarified the use of the command line option. For the second part of your comment, we did add some precisions on the specificity of the tools.
Initial text (L76-78): "The software can be used only with command lines making it compatible with high-performance computing (HPC) job schedulers."
Revised text (L125-130): "While the original goal in developing *mbctools* was providing assistance to novice users through its interface, it can also prove useful to bioinformaticians willing to integrate it as part of a wider data processing pipeline, since its series of mandatory operations (initial analysis described below) can be launched in a headless mode by feeding the program with a configuration file."
Text added to the revised manuscript (L130-134): "Additionally, addressing the challenge of processing large numbers of markers simultaneously across different kingdoms was one of the main motivations for designing *mbctools*, which allows for the simultaneous processing of multiple genetic markers and has a simplified graphical interface that guides the user through key parameters of the analysis"

6/ Figures 1-2-3: The quality of the figures appears to be quite low; I recommend using photos with greater resolution and better display. They appear to be not original creations, and designing in applications like Canva, POWERPOINT, or Inkscape can yield superior outcomes.

Author response: We agree with the reviewer's assessment, the quality of the mentioned figures are quite poor in the document. We added for Figure 1 the pdf version in additional files, for Figure 2, the Inkscape version (vertor pdf), and for Figure 3 as pdf version.

Prerequisites and Dependencies

The "Prerequisites and Dependencies" section provides a detailed overview of the technical requirements and setup needed to utilize mbctools efficiently. However, there are areas where further clarification and enhancement could improve readability and comprehension for potential users:

7/ Lines 93-100: This subsection clearly outlines the flexibility of mbctools in handling various data types and scenarios, such as single-end reads and paired-end reads. It may be beneficial to briefly mention examples of sequencing platforms (e.g., Illumina, Ion Torrent) that generate compatible data types, offering a direct reference point for researchers familiar with specific technologies.

Author response: We have added the suggested content to clarify the platform for under which VSEARCH operates.
Initial text (L94-95): "mbctools enables the analysis of amplicon data from multiple markers, obtained through NGS sequencing."
Revised text (L147-149): "*mbctools* enables the analysis of amplicon data from multiple markers, obtained from short reads (Illumina) or long reads (Oxford Nanopore Technology, PacBio)"

8/ Lines 101-107: The requirement for Python3.7 (or higher) and VSEARCH is straightforward, but it could be helpful to link directly to the VSEARCH GitHub page or documentation for users unfamiliar with this tool. Additionally, elaborating on the necessity of Powershell script execution for Windows users could aid in troubleshooting potential installation issues.

Author response: We think those are very good suggestions, and thus added details regarding those two points to mbctools' Github page (https://github.com/GuilhemSempere/mbctools?tab=readme-ov-file#requirements) and to the manuscript:
Initial text (L76-78): "The installation of Python3.7 (or higher) and VSEARCH (version 2.19.0 or higher) (14) is required to run the pipeline. Additionally, for Windows users, Powershell script execution must be enabled."
Revised text (L158-164): "The installation of Python3.7 (or higher) and VSEARCH (version 2.19.0 or higher, available at https://github.com/torognes/vsearch) is required to run the pipeline, and those binaries must be added to the PATH environment variable. Additionally, for Windows users, Powershell script execution must be enabled using the *Set-ExecutionPolicy* command, as the default Windows command prompt does not provide enough flexibility to run the software."

9/ Lines 108-132: The detailed setup for directory structure and necessary files is commendable for its thoroughness. However, this section could benefit from:
- Simplification: Breaking down this information into bullet points or a checklist might make the setup process seem less daunting.
- Visual Aids: Consider including a diagram or flowchart in addition to Figure 3 to visually represent the directory structure and file relationships.

Author response: While we appreciate the reviewer's feedback, we respectfully disagree. The section is already organized with bullet points L166-190 and Figure 3 eases the overview for the setup. Users can also refer to the Dataverse and Github page to visualize the contents of each file within the training set.

10/ Line 131: The instruction for launching mbctools is clear, but including information about common commands or operations that a new user might need to perform upon starting the tool could enhance usability. A brief "Getting Started" guide within this section, or as a reference to another part of the documentation, would be beneficial.

Author response: We added on Github a description for a more comprehensive way to start mbctools, under the section "Getting started": https://github.com/GuilhemSempere/mbctools?tab=readme-ov-file#getting-started
Additionally, as mentioned above, the need for Windows user to run the Set-ExecutionPolicy command is now put forward both in the GitHub documentation and in the manuscript (L162-163).

Software Features

11/ Lines 148-157: This passage effectively outlines the initial data analysis process. Clarifying whether the user has the flexibility to bypass the complete cycle for specific tasks or if they must follow a linear process would be helpful. This could address potential user questions about workflow customizability.

Author response: Thank you for pointing this out, we have developed the topic on workflow customization.
Initial text (L152-153): "Once those are completed, primers are removed, and the denoising process could be refined"
Revised text (L217-218): "Once those are completed, primers are removed, and the denoising process may be refined by tuning the cluster size threshold"
Text added to the revised manuscript (L224-226): "If, at any point, users wish to exit the program and conduct further analyses of their choice, mbctools allows it since it generates intermediate output files at each step."

12/ Lines 160-171: The detailed explanation of menu options and the flexibility to refine

analysis parameters is commendable. It might be useful to provide examples or case studies where adjusting these parameters significantly impacted the analysis outcome, offering practical insights into how users might leverage these features.(if exists)

Author response: We added the suggested content, explaining under which circumstances the parameters should be adjusted.
Text added to the revised manuscript (L239-244): "In most cases the default values for those parameters will work, but there might be situations where they need to be adapted, either globally (e.g., depending on the sequencing depth, the level of similarity between targeted genetic markers), at the marker (i.e., locus) level (e.g., according to its size), or at the sample level (e.g., to account for heterogeneous sample quality)"

Main pipeline description

13/ This section can be used as a supplementary file, rather than being included in the main text, or it can be uploaded to a github page as a document or even a video, which is more valuable to the user.

Author response: The “Main pipeline description” section has been moved to Annex 1 as supplementary material and we therefore changed the title of its original parent section from “Software features” to “Software features and interface overview”.

Conclusion

14/ Lines 398-402: The summary of mbctools’ functionalities is clear and concise. It might be beneficial to briefly recap the main advantages or unique features of mbctools compared to other available tools, reinforcing the reasons for its development and the gaps it aims to fill within the field.

Author response: : As suggested by the reviewer, we added clarity on the unique features of mbctools.
Text added to the revised manuscript (L335-338): "It is, to our knowledge, the only tool eliminating the need for the computer literacy normally required to utilize command-line based software as complex as the underlying VSEARCH, while allowing to process multiple genetic markers simultaneously"

15/ Lines 403-405: Mentioning the customization possibility for the taxonomic assignment step is crucial. It would be helpful to suggest some widely used tools or methods for this purpose, offering a starting point for users less familiar with the options available for taxonomic assignment.

Author response: Thank you for this suggestion, we added some examples of taxonomic affiliation approaches.
Text added to the revised manuscript (L341-345): "Although BLASTn is a commonly used and straightforward solution for this task, given the diversity of tools and approaches (k-mer based Kraken2 (31), Kaiju (32), phylogeny-based PhyloSift (33)), we leave to the user to proceed with taxonomic assignment, based on his own practice or the literature (18)"

16/ Lines 406-409: The transition towards mentioning the integration with metaXplor is smooth, but expanding on how mbctools specifically enhances data management, visualization, and accessibility through this compatibility could be enriching. Highlighting examples of how this feature has been utilized in previous studies or projects could provide concrete benefits.

Author response: we developed the topic by adding specific aspects and uses cases of metaXplor
Text added to the revised manuscript (L345-358): "To promote data management, visualization and long-term accessibility, mbctools offers a file conversion functionality compatible with the metaXplor web application which aims at centralizing online meta-omic data, while offering user-friendly means to interact with it. The metaXplor instance hosted at CIRAD (https://metaxplor.cirad.fr/) is indeed used by several teams to keep track of previous project data (be it private or public) and proves useful in helping scientists quickly recover precise information. As an example, metaXplor was utilized in the scope of 24 shotgun metagenomics projects based on VANA (Virion-Associated Nucleic Acids), conducted to uncover viral diversity (34). It allowed for the incremental building of a data repository that efficiently manages and provides means to analyze the extensive sequence datasets thus generated. This platform facilitated similarity-based searches and phylogenetic analyses, significantly enhancing the retrieval and re-analysis of data, thereby promoting viral discovery and classification"

17/ Future Directions and Development: While the note on ongoing development hints at mbctools' adaptability, a brief mention of specific areas or features under development could excite potential users about future updates. This could include mentioning planned improvements, new functionalities, or integration with other platforms and tools.

Author response: thank you for the suggestion, we developed this last part to highlight the future developpement of mbctools.
Initial text (L408-409): "mbctools is currently under development and can be adjusted to improve its applicability"
Revised text (L361-366): "This tool plays a central role in training sessions provided by our teams in developing countries, and its functionalities and user-friendliness are recurrently extended according to the feedback they generate. Depending on the success of this first release, future versions may add support for new sequencing technologies, embedded taxonomic assignment solutions or phylogenetic tree building"

2/ Review by Sourakhata Tirera, 17 Jun 2024 14:39

mbctools provides a wrapper that simplifies access to metabarcoding data analyses by providing an easy to install and well describe analysis steps along some options. It relies and depends only on vsearch software which users must download as an executable binary or install from source. It is also notable that mbctools is cross-plateform and can be used as command-line tool and even in HPC environments.

Despite this, I suggest authors improve some points.

1 Users must be informed that they won't get mbctools working unless they have accessible vsearch installation (via a PATH). This maybe stated in the software Readme section and pypi internet page.

Author response: We added to the GitHub repository documentation the description of a more comprehensive way to start mbctools, under the section "Getting started": https://github.com/GuilhemSempere/mbctools?tab=readme-ov-file#getting-started and provided more details in the manuscript:
Initial text (L76-78): "The installation of Python3.7 (or higher) and VSEARCH (version 2.19.0 or higher) (14) is required to run the pipeline. Additionally, for Windows users, Powershell script execution must be enabled."
Revised text (L158-164): "The installation of Python3.7 (or higher) and VSEARCH (version 2.19.0 or higher, available at https://github.com/torognes/vsearch) is required to run the pipeline, and those binaries must be added to the PATH environment variable. Additionally, for Windows users, Powershell script execution must be enabled using the *Set-ExecutionPolicy*

command, as the default Windows command prompt does not provide enough flexibility to run the software."

2 Windows users are often less used to command line and software installation procedures. I suggest a well detailed procedure specific to Windows OS.

Author response: We think this is a very good suggestion, and thus thoroughly added the suggested information to the Github page regarding Windows setup instructions: https://github.com/GuilhemSempere/mbctools?tab=readme-ov-file#procedure-for-windows-setup

3 Lines 99/100 : "The pipeline can accommodate a potentially unlimited number of samples". This is an overstatement. Even more the ability of the mbctools (depending on vsearch) to handle samples depends on the algorithmic complexity and the availability of computational resources. There is no estimation nor data on needed computational resources for processing, for example, the toy dataset. So, authors should this statement or remove it.

Author response: mbctools actually processes samples and loci sequentially (while leaving VSEARCH decide how to parallelize execution at each step) so requirements in terms of processing power should not grow with the number of samples of loci. These should only impact, obviously, the duration of the process and the required disk space to store output files. We do agree though that the phrasing needed to be nuanced, and changed it accordingly.
Initial text (L99-100): "The pipeline can accommodate a potentially unlimited number of samples and amplicons per sample"
Revised text (L153-155): "As it relies on the underlying VSEARCH software which is highly optimized in terms of efficiency and memory utilization, it is able to accommodate large datasets deriving from numerous samples and amplicons"

4 Line 227, there is a typo on « merge reads »

Author response: thank you, we have corrected the typo.

5 Line 297 for what the "[REF]" stands for? IF it is a reference, please provide it.

Author response: We have added the reference.

Overall, I suggest authors to lighten the software features section and add more results from real world or simulated datasets to highlight their tool's ability in diverse contexts. The

detailed version of the “software features” and usage can be then added to the git or any other public repository for future users.

Author response: We agree with the reviewer’s assessment, the Main pipeline description has been moved to Annex 1 as supplementary material and we changed the title of its original parent section from “Software features” to “Software features and interface overview”.