mbctools: A User-Friendly Metabarcoding and Cross-Platform Pipeline for Analyzing Multiple Amplicon Sequencing Data across a Large Diversity of Organisms

Christian Barnabé[a†], Guilhem Sempéré[bc†*], Vincent Manzanilla[a*], Joel Moo Millan[d], Antoine Amblard-Rambert[d] and Etienne Waleckx[ad]

[a]Institut de Recherche pour le Développement, Université de Montpellier, UMR INTERTRYP IRD, CIRAD, University of Montpellier, France.

[b]Centre de Coopération Internationale en Recherche Agronomique pour le Développement, UMR INTERTRYP IRD, CIRAD, Campus international de Baillarguet, Montpellier, France

[c]South Green Bioinformatics Platform, Biodiversity, Montpellier, France

[d]Laboratorio de Parasitología, Centro de Investigaciones Regionales "Dr Hideyo Noguchi", Universidad Autónoma de Yucatán, Mérida, México

†Co–first authors, contributed equally

*Corresponding authors: guilhem.sempere@cirad.fr, vincent.manzanilla@ird.fr

Abstract

We developed a python package called *mbctools*, designed to offer a cross-platform tool for processing amplicon data from various organisms in the context of metabarcoding studies. It can handle the most common tasks in metabarcoding pipelines such as paired-end merging, primer trimming, quality filtering, sequence denoising, zero-radius operational taxonomic unit (ZOTU) filtering, and has the capability to process multiple genetic markers simultaneously. *mbctools* is a menu-driven program that eliminates the need for expertise in command-line skills and ensures documentation of each analysis for reproducibility purposes. 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. 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. To adapt to the diversity of projects in metabarcoding, we facilitate the reprocessing of datasets with the possibility to adjust parameters. *mbctools* can also be launched in a headless mode, making it suited for integration into pipelines running on High-Performance Computing environments. *mbctools* is available at https://github.com/GuilhemSempere/mbctools, https://pypi.org/project/mbctools/.

Keywords: metabarcoding, amplicon, reproducibility, VSEARCH, pipeline

Introduction

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).

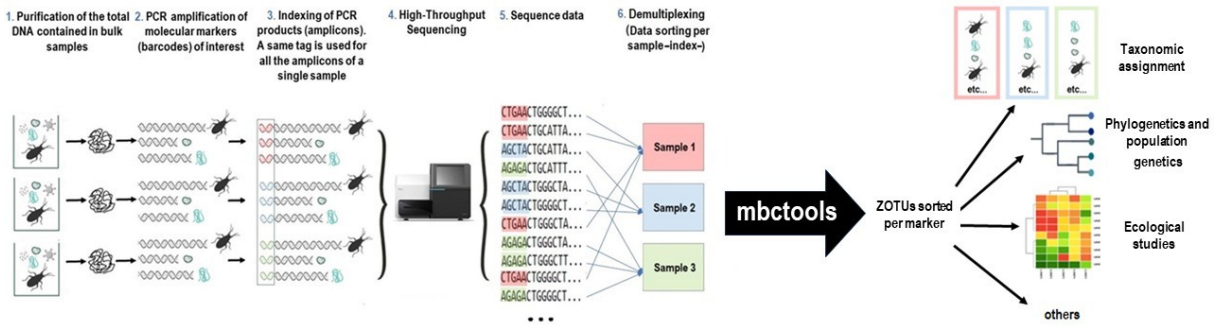


Figure 1: Metabarcoding workflow example and area of operation of mbctools

The increasing demand for efficient and versatile pipelines to analyze amplicon data has led to the development of over 30 programs and pipelines in recent years (17). However, these applications often present some challenges in terms of reproducibility and user-friendliness for non-expert users in the field of bioinformatics. 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. This accessibility ensures that non-expert users can effectively and autonomously analyze their data without experiencing difficulties in reproducibility or requiring extensive bioinformatics knowledge.

Among pipelines and algorithms used in the field of metabarcoding data processing, substantial variations in sensitivity and specificity are observed (18,19). From these benchmarking studies, VSEARCH appears appropriate for our applications. mbctools' main purpose is to facilitate the use of VSEARCH, which has demonstrated strong performance in handling large-scale next-generation sequencing (NGS) data (20), and it thus could be considered a "wrapper", hiding the complexity of command-line to the users. It thus allows them to easily process sequencing reads and obtain meaningful outputs, i.e, zero-radius operational taxonomic units (ZOTUs (21), almost equivalent to ASVs (amplicon sequence variants)) along with abundance information.

mbctools processes demultiplexed metabarcoding data obtained from one or multiple genetic markers to distribute cleaned and filtered sequences into ZOTUs sorted per marker, which can be used for taxonomic assignment or further analyzes such as phylogeny (Figures 1, 2). 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. 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.

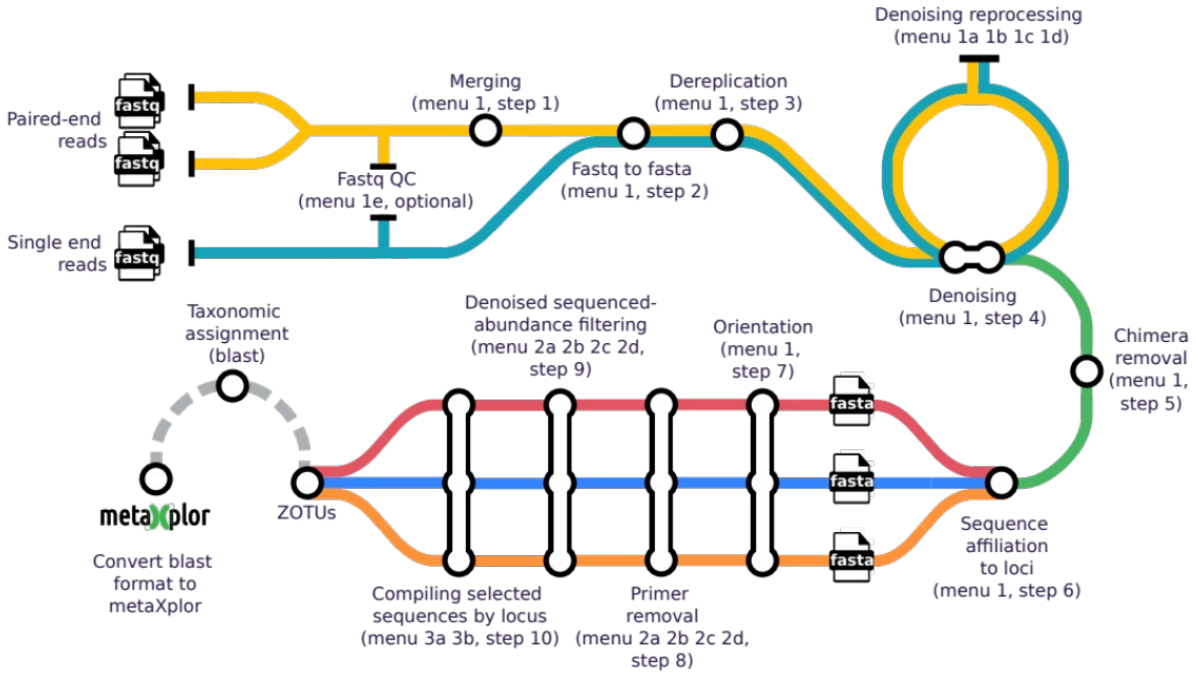


Figure 2: Overview of the mbctools workflow, where each station icon represents a distinct step in the process. In the initial analysis (menu 1), data processing begins with the merging for paired-end reads (step 1) and the conversion of reads from fastq to fasta (step 2). This is followed by dereplication (step 3), denoising (step 4, which can be re-run with various options via menu 1a, 1b, 1c, 1d), chimera removal (step 5), sequence affiliation to markers (step 6), and finally, sequence orientation harmonization (step 7). Following this, primer removal (step 8) and denoised sequence abundance filtering (step 9) are applied. The last step in the analysis involves obtaining per-marker fasta files containing retained ZOTUs (step 10). The menu labeled as "4" comes into play after taxonomic assignment, tuning outputs for visualization in metaXplor and ensuring long-term data accessibility.

Prerequisites and dependencies

Input data types - sequencing tech & lab

mbctools enables the analysis of amplicon data from multiple markers, obtained from short reads (Illumina) or long reads (Oxford Nanopore Technology, PacBio). It can handle both single-end reads and paired-end reads, provided that the latter can be merged. In the eventuality where the merging of paired-end reads is prevented by either large amplicon lengths (leading to insufficient overlap zone) or the presence of low complexity sequences, *mbctools* is able to utilize only forward reads for analysis. 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.

Running environment

mbctools is a versatile cross-platform python package running seamlessly on Windows, Unix, or OSX systems. 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. The *mbctools* package, available from the Python Package Index (PyPI) (https://pypi.org/project/mbctools/), can be installed and upgraded using the command line tool pip (*pip3 install mbctools*).

Before starting the analyses, the package requires mandatory directories and files to be pre-arranged as illustrated in Figure 3:

- A default directory, named 'fastq', should contain all the fastq files corresponding to the R1 (and R2, when working with paired-ends sequencing) reads. These files are expected to be demultiplexed, meaning that each fastq file should contain reads from a single sample.
- A default directory named 'refs' that contains one reference file per genetic marker in the multi-FASTA format. Each of them should contain one or more unaligned reference sequences, from the same strand and without gaps. Providing multiple reference sequences for a given marker improves diversity support and thus leads to a better assignment accuracy. In addition, the 'sequence orientation fixing' step runs faster when using several reference sequences. Filenames must end with the extension '.fas' and must be called by the name of the corresponding marker, referred to as locus in *mbctools* (Figure 3).
- 'lociPE.txt' provides the list of markers (one per line) corresponding to paired-end R1/R2 reads.
- 'lociSE.txt' provides the list of markers (one per line) corresponding to single-end reads and for which only R1 reads will be analyzed.
- 'samples.txt' includes a list of sample names, where each name appears on a separate line.
- 'fastqR1.txt' contains the names of the *fastq* R1 files corresponding to the samples.
- 'fastqR2.txt' contains the names of the *fastq* R2 files corresponding to the samples.

The contents of the last three files must be sorted in the same order, i.e., line (i) of 'samples.txt' shall correspond to line (i) of both 'fastqR1.txt' and 'fastqR2.txt'.

mbctools shall be launched from within the working directory in a shell console, typically by typing 'mbctools' if installed as a package.

```
/YOUR_PROJECT_PATH/
├────fastq
│  └────sample1.R1.fastq
│  └────sample1.R2.fastq
│  └────sample2.R1.fastq
│  └────sample2.R2.fastq
│  └────...
├────refs
│  └────locus1.fas
│  └────locus2.fas
│  └────...
├────samples.txt
├────fastqR1.txt
├────fastqR2.txt
├────lociPE.txt
├────lociSE.txt
```

Figure 3: Pre-required files and directories

Software features and interface overview

mbctools features a series of menus, each containing two to four options as illustrated in Figures 4 to 8. On startup, the program displays the main menu shown in Figure 4. This menu presents descriptive keywords that enable users to easily navigate through the available menus and sections. From each menu, the user is required to select one single option at a time.

```
------------------------------ mbctools v1.0-beta2 - MAIN MENU ------------------------------

Please cite this software as follows:
mbctools:       A User-Friendly Metabarcoding and Cross-Platform Pipeline for Analyzing
                Multiple Amplicon Sequencing Data across a Large Diversity of Organisms
C. Barnabé, G. Sempéré, V. Manzanilla, J. Moo-Millan, A. Amblard-Rambert and E. Waleckx
bioRxiv, 2024.02.08.579441, https://doi.org/10.1101/2024.02.08.579441


NAVIGATION CONVENTIONS:
Entering 'back' returns to the program upper level, if any
Entering 'home' returns to this main menu
Entering 'exit' leaves the program
Validating without typing anything applies the default value, if any


We recommend executing procedures in the provided order:

1 -> INITIAL ANALYSIS (mandatory): read merging, sample-level dereplication, sequence clustering,
        chimera detection, affiliation of sequences to loci, and sequence re-orientation

2 -> PRIMER REMOVAL AND SELECTION OF MINIMUM SEQUENCE ABUNDANCE LEVELS ACCORDING TO USER-DEFINED THRESHOLDS

3 -> GENERATION OF A UNIQUE SEQUENCE FILE FOR EACH LOCUS (comprising all samples' data)

4 -> EXPORTING ANALYSIS RESULTS INTO metaXplor FORMAT
```

Figure 4: Contents of the main menu, explaining navigation conventions and providing access to the four main sections.

To analyze a new dataset, it is necessary to perform a complete initial

analysis by selecting the first option from both the first and subsequent menu. The *initial analysis* processes the data through a succession of steps such as merging, sample-level dereplication, sequence denoising, chimera detection and removal, affiliation of sequences to markers (e.g., loci), and sequence re-orientation (menu 1, Figure 5). Once those are completed, primers are removed, and the denoising process may be refined by tuning the cluster size threshold (menu 2, Figure 6). Finally, centroid sequences associated with each marker can be exported to fasta files (menu 3, Figure 7). The above represents the complete cycle of data processing. Additionally (menu 4, Figure 8), *mbctools* software offers the possibility to export results in a format that is compatible with the metaXplor platform (22). The entire process is described as supplementary material.

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.

```
--- MENU 1: BASIC ANALYSIS - only option 1 is strictly mandatory ---

1  -> INITIAL ANALYSIS (mandatory): read merging, sample-level dereplication, sequence clustering,
        chimera detection, affiliation of sequences to loci, and sequence re-orientation

1a -> Re-analyze all loci, from the clustering step, modifying parameters

1b -> Re-analyze only one locus of paired-end amplicon (merged reads), modifying parameters

1c -> Re-analyze only one locus of single-end amplicon (R1 only), modifying parameters

1d -> Re-analyse a given sample, modifying parameters

1e -> Optional quality checking of fastq files (slow)
```

Figure 5: Contents of section 1 menu, in which only option 1, INITIAL ANALYSIS, is mandatory as it executes all steps required to proceed with menu 2. Options 1a to 1d may be used after option 1 in case the user wants to refine some parameters. Analysis 1e is optional and provides information about the reads' quality.

When option 1 (initial analysis) is selected from menu 1, the menu shown on Figure 5 allows to run the initial analysis (section 1), namely read merging, sample-level dereplication, sequence clustering (denoising), chimera detection, affiliation of sequences to markers, and sequence re-orientation. In the same menu, four other sections (1a, 1b, 1c, 1d) provide means to re-process some particular markers and/or samples with adjusted parameters. 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). Option 1e may be used to assess the read quality for each sample, and write results to files with suffix "*.quality.txt" in the outputs directory.

When the initial analysis is completed, menu 2 should be used to remove primer sequences, and to exclude denoised sequences with an abundance considered too low for being relevant (Figure 6). This operation may be launched either in batch mode, or sample by sample in case different thresholds need to be

applied.

```
--- MENU 2: PRIMER REMOVAL, SELECTION OF MINIMUM SEQUENCE ABUNDANCE LEVELS ACCORDING TO USER-DEFINED THRESHOLDS ---

2a -> Apply the SAME size threshold for ALL SAMPLES for the loci based on PAIRED-END reads (R1/R2 merged)
        i.e. you want to keep only sequences whose abundance is greater than x% of the total number of
        sequences for a given sample. This threshold of x% can be chosen for each locus.

2b -> Apply the SAME size threshold for ALL SAMPLES for the loci based on SINGLE-END reads (R1 only)
        same as option 2a but only using the R1 reads instead of merged ones.

2c -> Apply a SPECIFIC size threshold for a given sample, for the loci based on PAIRED-END reads (R1/R2 merged)
        i.e. you want to modulate the threshold of x% by locus but also by sample within a particular locus.

2d -> Apply a SPECIFIC size threshold for a given sample, for the loci based on SINGLE-END reads (R1 only)
        same as option 2c but only using the R1 sequences instead of merged ones.
```

Figure 6: Menu 2 combines two procedures that need to be launched successively for each marker: primer removal and application of an abundance threshold for retaining centroid sequences.

In menu 3, the program groups sample files for each marker, creating a new file per locus with the suffix "*_allseq_select.fasta" (Figure 7). During this process, mbctools can, as an option, dereplicate sequences at the project level, ensuring that each entry in the fasta file is unique. In such cases, an additional tabulated file is generated to maintain a record of sequence abundance across the samples. Optionally generated files are then named with the following suffixes: "*_allseq_select_derep.fasta", "*_allseq_select_derep.tsv".

```
--- MENU 3: GENERATION OF A UNIQUE SEQUENCE FILE FOR EACH LOCUS (comprising all samples' data) ---

3a -> Process all loci at once

3b -> Process loci one by one
```

Figure 7: Contents of section 3 menu, allowing to obtain per-marker (per-locus) fasta files. Menu 3 provides means to generate fasta files containing the retained sequences, either for all markers (loci) at once, or for a given marker.

Finally, menu 4 converts *mbctools* outputs into a format suitable for importing into metaXplor (Figure 8), a web-oriented platform for storing, sharing, exploring and manipulating metagenomic pipeline outputs. metaXplor significantly contributes to enhancing the FAIR (Findable, Accessible, Interoperable, and Reusable) (23) characteristics of this specific data type by serving as a robust solution for centralized data management, intuitive visualization, and long-term accessibility.

```
--- MENU 4: EXPORTING ANALYSIS RESULTS INTO metaXplor FORMAT ---

4a -> Generate sequence files
        Compiles all sequences selected for all loci into a single fasta
        Outputs a .tsv file indicating samples weights for each sequence

4b -> Generate assignment file
        Converts blastn results (obtained from blasting above-mentioned fasta file) from 'Hit table (text)'
        (format #7) into metaXplor format

4c -> Build metaXplor-format sample metadata file from provided tabulated file

4d -> Compresses all metaXplor files into a final, ready to import, zip archive
```

Figure 8: Contents of section 4 menu, guiding users in obtaining necessary files to create a metaXplor import archive (i.e., sample metadata, sequence fasta and abundance, assignment to reference taxa) in order to summarize all project results.

Test data

The test data were generated in the scope of a study conducted on Chagas disease, a zoonotic parasitic illness caused by *Trypanosoma cruzi*. This disease is primarily transmitted to mammals through the feces and urine of hematophageous insects known as triatomines, commonly referred to as kissing bugs (Hemiptera: Reduviidae). *Triatoma dimidiata* is the main vector in the Yucatan Peninsula in Mexico. In 2019, the owner of an isolated rural dwelling collected thirty-three samples of *T. dimidiata* individuals from a conserved sylvatic site within the Balam Kú State Reserve, located in Campeche, Mexico (Figure 9). We used a metabarcoding approach to simultaneously characterize the genetic diversity of the vector and identify its blood meals, characterize its hindgut and midgut microbiome composition, and assess the genetic diversity of *T. cruz*i infecting *T. dimidiata* in this sylvatic ecotope. We extracted the total DNA from each *T. dimidiata* gut, and amplified different selected markers of interest for our question: (i) for *T. dimidiata* genotyping, a fragment of the vector's Internal Transcribed Spacer 2 (ITS-2) nuclear marker was amplified with primers ITS2_200F (5’-TCG YAT CTA GGC ATT GTC TG-3’) and ITS2_200R (5-’CTC GCA GCT ACT AAG GGA ATC C-3’) previously described in (24); (ii) for the identification of vertebrate blood meals, a fragment of the vertebrate 12S rRNA gene was amplified with primers L1085 (5’- CCC AAA CTG GGA TTA GAT ACC C-3’) and H1259 (5’- GTT TGC TGA AGA TGG CGG TA-3’) (25,26); (iii) for the identification of hindgut and midgut bacterial microbiome composition, a fragment of the bacterial 16S rRNA gene was amplified with primers E786F (5’-GAT TAG ATA CCC TGG TAG-3’) and U926R (5'-CCG TCA ATT CCT TTR AGT TT-3') (27); and (iv) to accurately genotype *T. cruzi*, we targeted markers COII, GPI, ND1 and the intergenic region of the mini-exon gene, using the primers described in (10). Then we pooled the different amplicons obtained per *T. dimidiata* specimen and sequenced them on an Illumina MiSeq. On average, we obtained 222,244 reads per sample. The reads were processed with the *mbctools* package, whose outputs were imported into metaXplor. In *mbctools*, we retained the reads with a minimum length of 100bp, clusters with a minimum size of 8 and used an alpha value of 2. The taxonomic identification of the ZOTUs was performed using

BLASTn (28). The exported CSV file with the metadata served as input for metaXplor. The analysis results are directly available on the web application under the dataset 'mbctools_review' (29) (Figure 9).

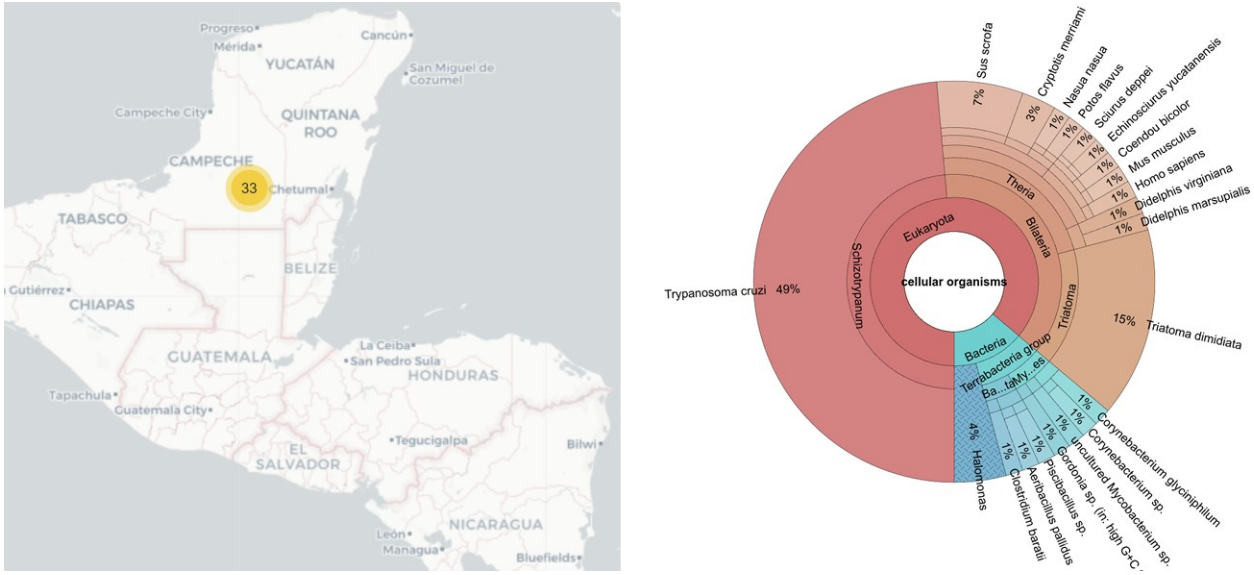


Figure 9: Graphical output from the metaXplor web interface. On the left side the geographic origin of the thirty three samples. On the right side the taxonomic proportion of each identified organism

We identified an average of 33 ZOTUs per *T. dimidiata* sample, and only 9.61% of the reads were discarded during the analysis (Table 1). After analyses we found thirteen blood-meal sources (the most frequently identified being *Homo sapiens*, *Sus scrofa* and *Mus musculus*), six bacterial Orders (Bacillales, Oceanospirillales and Micrococcales being the most abundant), along with *T. cruzi* and *T. dimidiata* (Figure 9). The test data input and output files are freely accessible on the DataSuds repository (30).

Sample	Reads per sample	Merging	Number of clusters	clusters with no chimera	Cluster per marker							
					12S	16S	COII	GPI	ITS2	ND1	MINEX	MINEX_SE
CON001	215137	0.8696	11	11	0	15	0	0	2	3	6	0
CON002	268415	0.9229	17	13	18	12	0	0	2	0	0	0
CON003	242823	0.8636	87	87	0	9	0	1	2	0	12	0
CON004	205358	0.9344	3	3	0	11	0	0	3	0	0	0
CON005	149945	0.9408	7	7	1	6	0	0	2	0	0	0
CON006	180763	0.9311	2	2	0	11	0	0	3	0	0	0
CON007	7898	0.891	2	2	1	0	0	0	2	0	0	0
CON008	200589	0.924	3	3	0	19	0	0	2	0	0	0
CON009	261914	0.9226	2	2	0	15	0	0	2	0	0	0
CON010	283721	0.9325	139	126	16	16	0	0	2	0	0	0
CON011	186936	0.9112	1	1	0	16	0	0	2	0	0	0
CON012	194776	0.8326	61	60	14	10	2	1	2	0	24	4
CON013	189441	0.9303	14	14	4	8	0	0	4	0	5	0
CON014	151880	0.9086	25	25	0	8	0	0	30	0	0	0
CON015	221302	0.9174	1	1	0	14	0	0	2	0	0	0
CON016	139546	0.8699	46	46	2	6	2	1	10	0	14	3
CON017	228038	0.8812	3	3	0	11	0	0	2	0	0	0
CON018	258481	0.9012	28	27	6	12	0	0	15	0	0	0
CON019	251403	0.7904	77	77	1	14	4	2	10	0	3	0
CON020	223972	0.9283	24	24	2	1	0	0	4	0	0	0
CON021	309618	0.9346	8	8	0	15	0	0	14	0	0	0
CON022	197125	0.8479	83	83	11	8	2	2	4	0	5	0
CON023	193857	0.7558	53	53	1	7	2	1	5	5	15	0
CON024	190310	0.9132	62	62	13	20	0	0	2	0	0	0
CON025	273603	0.9104	44	37	29	15	0	0	2	0	0	0
CON026	284204	0.945	2	2	1	17	0	0	2	0	0	0
CON027	436822	0.9225	5	5	1	34	0	0	2	0	0	0
CON028	243235	0.9385	63	63	1	12	0	0	4	0	0	0
CON029	219907	0.9486	13	13	4	11	0	0	2	0	0	0
CON030	306738	0.9456	17	17	0	10	0	0	26	0	0	0
CON031	172815	0.9398	87	87	1	11	0	0	2	0	0	0
CON032	206078	0.9069	34	33	1	13	0	0	24	0	0	0
CON033	237429	0.9158	92	92	37	14	0	0	2	0	0	0
Average	222244.8	0.9	33.8	33	5	12.2	0.36	0.24	5.88	0.24	2.55	0.21

Table 1: Statistics concerning the processing of the test data set with the number of reads per sample and the attribution to each marker.

Conclusion

mbctools facilitates the processing of amplicon datasets for metabarcoding studies by offering menus that guide users through the various steps of the data analysis. 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. *mbctools* offers the possibility to reprocess specific markers and samples with customized parameters, providing means to adapt to the diversity of metabarcoding datasets. In the end, it generates *fasta* files composed of the filtered ZOTUs ready for the taxonomic assignment step. 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). 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. Although we initially developed the *mbctools* pipeline and software for reproducibility purposes across our teams working on infectious diseases, the pipeline's strength resides in its capacity to handle diverse data and be tailored to different research needs. 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.

Authors' contributions

C.B and G.S designed and conceived the software based on E.W.'s specifications. V.M. did the Python packaging. E.W, G.S, V.M, A.A-R and C.B wrote the article. A.A-R. provided the data to test mbctools. All authors read and approved the final manuscript.

Funding

This work received financial support from CONACYT (National Council of Science and Technology, Mexico) Basic Science (Project ID: CB2015-258752) and National Problems (Project ID: PN2015-893) grants awarded to EW, as well as from IRD (French National Research Institute for Sustainable Development).

Data, scripts, code, and supplementary information availability

mbctools is openly accessible on Github at https://github.com/GuilhemSempere/mbctools and on PyPI at https://pypi.org/project/mbctools/. The main pipeline description is available as supplementary material, as are higher-definition versions of Figures 1 to 3.

Acknowledgements

The authors would like to thank Frédéric Mahé for taking the time to introduce

VSEARCH functionalities, as well as Dorian Grasset for testing the OSX version and creating a GitHub action to automatize PyPI package generation.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

References

1. TABERLET P, COISSAC E, POMPANON F, BROCHMANN C, WILLERSLEV E. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology. 2012;21(8):2045-50.

2. Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proc R Soc Lond B. 7 févr 2003;270(1512):313-21.

3. Holdaway RJ, Wood JR, Dickie IA, Orwin KH, Bellingham PJ, Richardson SJ, et al. Using DNA metabarcoding to assess New Zealand's terrestrial biodiversity. New Zealand Journal of Ecology. 2017;41(2):251-62.

4. Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E. ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Research. 1 nov 2011;39(21):e145-e145.

5. Hebert PD, Ratnasingham S, Zakharov EV, Telfer AC, Levesque-Beaudin V, Milton MA, et al. Counting animal species with DNA barcodes: Canadian insects. Philosophical Transactions of the Royal Society B: Biological Sciences. 2016;371(1702):20150333.

6. CBOL Plant Working Group 1, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, et al. A DNA barcode for land plants. Proceedings of the National Academy of Sciences. 2009;106(31):12794-7.

7. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the national academy of Sciences. 2012;109(16):6241-6.

8. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the national academy of sciences. 2011;108(supplement_1):4516-22.

9. Lauthier JJ, Tomasini N, Barnabé C, Rumi MMM, D'Amato AMA, Ragone PG, et al. Candidate targets for Multilocus Sequence Typing of Trypanosoma cruzi: validation using parasite stocks from the Chaco Region and a set of reference strains. Infection, genetics and evolution. 2012;12(2):350-8.

10. Jaime LD, Aracely LM, Paulina OM, Dumonteil E, Barnabé C, Waleckx E, et al. Molecular characterization of four Mexican isolates of Trypanosoma cruzi and their profile

susceptibility to nifurtimox. Acta parasitologica. 2022;67(4):1584-93.

11. Barnabé C, Brenière SF, Santillán-Guayasamín S, Douzery EJP, Waleckx E. Revisiting gene typing and phylogeny of Trypanosoma cruzi reference strains: Comparison of the relevance of mitochondrial DNA, single-copy nuclear DNA, and the intergenic region of mini-exon gene. Infection, Genetics and Evolution. nov 2023;115:105504.

12. Hernández-Andrade A, Moo-Millan J, Cigarroa-Toledo N, Ramos-Ligonio A, Herrera C, Bucheton B, et al. Metabarcoding: a powerful yet still underestimated approach for the comprehensive study of vector-borne pathogen transmission cycles and their dynamics. In: Vector-Borne Diseases-Recent Developments in Epidemiology and Control. IntechOpen; 2019.

13. Leray M, Knowlton N. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Sciences. 2015;112(7):2076-81.

14. Thomsen PF, Willerslev E. Environmental DNA–An emerging tool in conservation for monitoring past and present biodiversity. Biological conservation. 2015;183:4-18.

15. Raclariu AC, Paltinean R, Vlase L, Labarre A, Manzanilla V, Ichim MC, et al. Comparative authentication of Hypericum perforatum herbal products using DNA metabarcoding, TLC and HPLC-MS. Sci Rep. 2 mai 2017;7(1):1291.

16. Dumonteil E, Ramirez-Sierra MJ, Pérez-Carrillo S, Teh-Poot C, Herrera C, Gourbière S, et al. Detailed ecological associations of triatomines revealed by metabarcoding and next-generation sequencing: implications for triatomine behavior and Trypanosoma cruzi transmission cycles. Scientific reports. 2018;8(1):4140.

17. Hakimzadeh A, Asbun AA, Albanese D, Bernard M, Buchner D, Callahan B, et al. A pile of pipelines: an overview of the bioinformatics software for metabarcoding data analyses. 2023;

18. Bailet B, Apothéloz-Perret-Gentil L, Baričević A, Chonova T, Franc A, Frigerio JM, et al. Diatom DNA metabarcoding for ecological assessment: Comparison among bioinformatics pipelines used in six European countries reveals the need for standardization. Science of the Total Environment. 2020;745:140948.

19. Mathon L, Valentini A, Guérin P, Normandeau E, Noel C, Lionnet C, et al. Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. Molecular Ecology Resources. 2021;21(7):2565-79.

20. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.

21. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv. 2016;081257.

22. Sempéré G, Pétel A, Abbé M, Lefeuvre P, Roumagnac P, Mahé F, et al. metaXplor: an interactive viral and microbial metagenomic data manager. GigaScience. 29 janv 2021;10(2):giab001.

23. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3(1):1-9.

24. Richards B, De La Rua NM, Monroy C, Stevens L, Dorn PL. Novel polymerase chain reaction-restriction fragment length polymorphism assay to determine internal transcribed

spacer-2 group in the Chagas disease vector, Triatoma dimidiata (Latreille, 1811). Mem Inst Oswaldo Cruz. juin 2013;108(4):395-8.

25. Kitano T, Umetsu K, Tian W, Osawa M. Two universal primer sets for species identification among vertebrates. Int J Legal Med. 21 août 2007;121(5):423-7.

26. Moo-Millan JI, Arnal A, Pérez-Carrillo S, Hernandez-Andrade A, Ramírez-Sierra MJ, Rosado-Vallado M, et al. Disentangling Trypanosoma cruzi transmission cycle dynamics through the identification of blood meal sources of natural populations of Triatoma dimidiata in Yucatán, Mexico. Parasites Vectors. déc 2019;12(1):572.

27. Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. Journal of Microbiological Methods. déc 2003;55(3):541-55.

28. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic acids research. 2008;36(suppl_2):W5-9.

29. metaXplor: an interactive viral and microbial metagenomic data manager [Internet]. 2024 [cité 5 févr 2024]. metaXplor: mbctools data set. Disponible sur: https://metaxplor.cirad.fr/metaXplor/main.jsp?module=mbctools_review

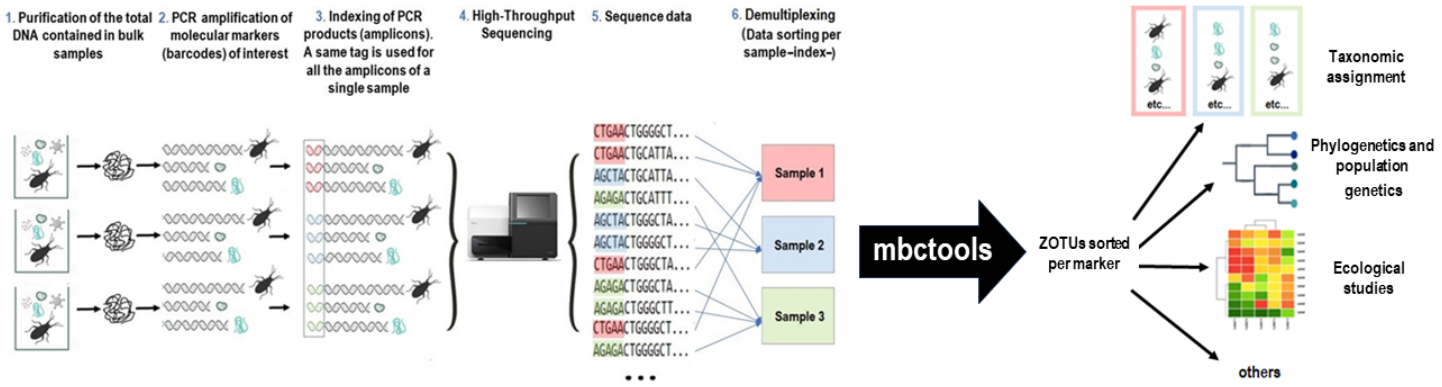
30. Manzanilla V, Barnabé C, Guilhem S, Millan JM, Waleckx E. mbctools: A User-Friendly Metabarcoding and Cross-Platform Pipeline for Analyzing Multiple Amplicon Sequencing Data across a Large Diversity of Organisms [Internet]. DataSuds; 2024 [cité 3 sept 2024]. Disponible sur: https://dataverse.ird.fr/citation?persistentId=doi:10.23708/W3TODJ

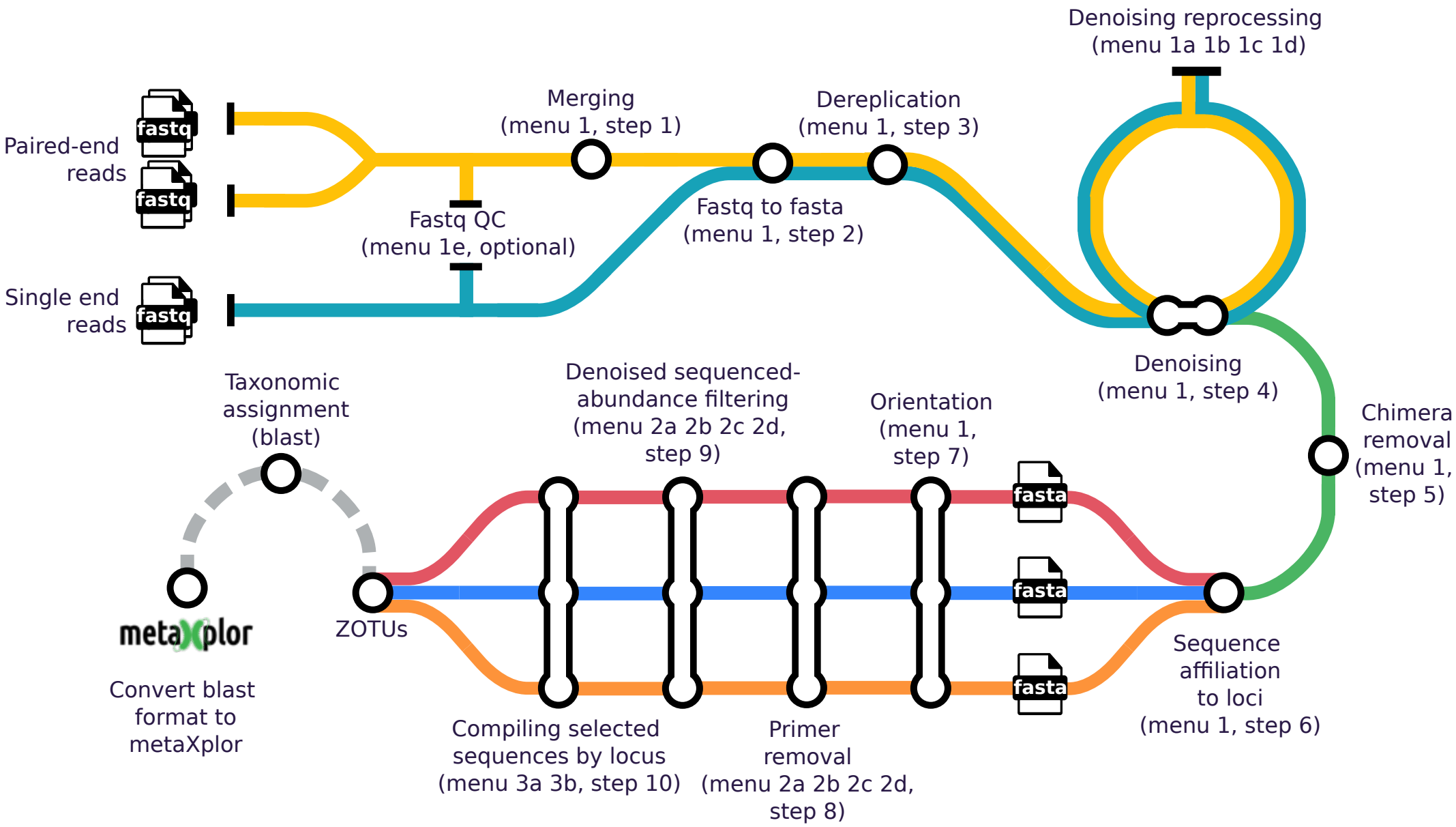
31. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome biology. 2019;20:1-13.

32. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature communications. 2016;7(1):11257.

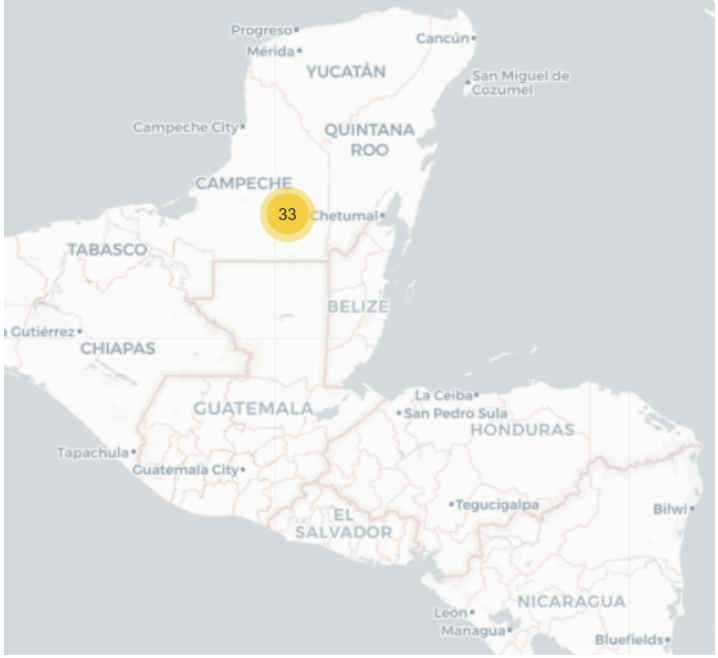
33. Darling AE, Jospin G, Lowe E, Matsen IV FA, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;2:e243.

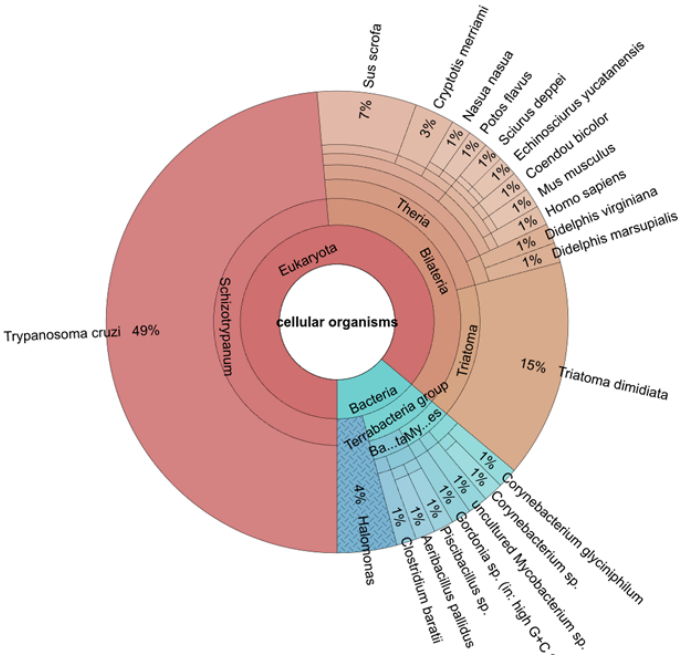
34. Moubset O, François S, Maclot F, Palanga E, Julian C, Claude L, et al. Virion-Associated Nucleic Acid-Based Metagenomics: A Decade of Advances in Molecular Characterization of Plant Viruses. Phytopathology®. nov 2022;112(11):2253-72.


1. Purification of the total DNA contained in bulk samples
2. PCR amplification of molecular markers (barcodes) of interest
3. Indexing of PCR products (amplicons). A same tag is used for all the amplicons of a single sample
4. High-Throughput Sequencing
5. Sequence data
6. Demultiplexing (Data sorting per sample-index-)
CTGAACTGGGGCT...
CTGAACTGCATTA...
AGCTACTGCATTA...
AGAGACTGCATTT...
AGCTACTGGGCTA...
AGCTACTGGGGCT...
CTGAACTGGGCTA...
AGAGACTGGGCTA...
AGAGACTGGGCTT...
CTGAACTGGGGCT...
AGAGACTGGGGCT...
Sample 1
Sample 2
Sample 3
mbctools
ZOTUs sorted per marker
etc...
etc...
etc...
Taxonomic assignment
Phylogenetics and population genetics
Ecological studies
others


Denoising reprocessing
(menu 1a 1b 1c 1d)
Paired-end
reads
fastq
fastq
Merging
(menu 1, step 1)
Dereplication
(menu 1, step 3)
Fastq QC
(menu 1e, optional)
Fastq to fasta
(menu 1, step 2)
Single end
reads
fastq
Denoising
(menu 1, step 4)
Chimera
removal
(menu 1,
step 5)
Taxonomic
assignment
(blast)
Denoised sequenced-
abundance filtering
(menu 2a 2b 2c 2d,
step 9)
Orientation
(menu 1,
step 7)
fasta
fasta
fasta
metaXplor
ZOTUs
Sequence
affiliation
to loci
(menu 1, step 6)
Convert blast
format to
metaXplor
Compiling selected
sequences by locus
(menu 3a 3b, step 10)
Primer
removal
(menu 2a 2b 2c 2d,
step 8)

```
/YOUR_PROJECT_PATH/
├────fastq
│   └────sample1.R1.fastq
│   └────sample1.R2.fastq
│   └────sample2.R1.fastq
│   └────sample2.R2.fastq
│   └────...
├────refs
│   └────locus1.fas
│   └────locus2.fas
│   └────...
├────samples.txt
├────fastqR1.txt
├────fastqR2.txt
├────lociPE.txt
├────lociSE.txt
```


Progreso
Mérida
Cancún
YUCATÁN
San Miguel de Cozumel
Campeche City
QUINTANA ROO
CAMPECHE
33
Chetumal
TABASCO
BELIZE
Cutiérrez
CHIAPAS
GUATEMALA
La Ceiba
San Pedro Sula
HONDURAS
Tapachula
Guatemala City
Tegucigalpa
Bilwi
EL SALVADOR
NICARAGUA
León
Managua
Bluefields


cellular organisms
Eukanyota
Bacteria
Schizotrypanum
Trypanosoma cruzi 49%
Theria
Bilateria
Triatoma
Terrabacteria group
Ba...taMy...es
Halomonas
7% Sus scrofa
3% Cryptotis merriami
1% Nasua nasua
1% Potos flavus
1% Sciurus deppei
1% Echinosciurus yucatanensis
1% Coendou bicolor
1% Mus musculus
1% Homo sapiens
1% Didelphis virginiana
1% Didelphis marsupialis
15% Triatoma dimidiata
1% Corynebacterium glyciniphilum
1% Corynebacterium sp.
1% uncultured Mycobacterium sp.
1% Gordonia sp. (in: high G+C
1% Piscibacillus sp.
1% Aeribacillus pallidus
1% Clostridium baratii
4%