An R Package for Streamlined 16S rRNA Microbiome Analysis

Dr. Lourenco's Lab, University of Georgia

Initial Release: March 2025

mbX

The **mbX** package is a comprehensive, user-friendly toolkit designed to streamline the 16S rRNA gene microbiome data analysis pipeline, particularly after taxonomic classification has been concluded. Whether you're a bioinformatician looking to automate your workflow or a researcher eager to generate reproducible, publication-quality outputs, mbX provides very useful solutions—from taxonomic data cleaning to dynamic visualization and stats.

mbX is crafted to handle the complexities of microbial data with minimal manual intervention. It is particularly geared toward users who work with raw taxonomic data and a corresponding metadata, such as those generated by QIIME2. The package consists of core functions—most notably **ezclean** and **ezviz**—that automate:

- **Taxonomic Data Cleaning:**
  The **ezclean** function transforms raw taxonomic data into a standardized (converted to relative abundance), structured format ready for downstream processing. It integrates seamlessly with metadata files and works flawlessly with outputs from popular platforms like QIIME2.

- **Data Visualization:**
  The **ezviz** function builds upon the cleaned data by further processing, grouping, averaging, and aggregating taxa. It then produces high-resolution, publication-ready stacked bar plots that depict the relative abundance of microbial taxa at the desired taxonomic level. Users can choose to retain either the top abundant taxa or filter based on a user-defined threshold, ensuring that the final visualization is both clear and informative.

- **(under development) Statistical Analysis (ezstat):**
  A future addition to **mbX** will offer streamlined statistical analyses, enabling users to derive even deeper insights from their microbiome datasets—all within the same cohesive workflow—producing publication-ready statistical outputs such as group averages and p-values.

- **Workflow Automation and Reproducibility:**
  mbX is designed to ensure reproducible analysis by automatically generating consistent output file names, cleaning up intermediate files, and providing clear console messages throughout the process. This structured approach minimizes errors and keeps your working directory organized.

Installation and Getting Started

Install the **mbX** package from CRAN (or your preferred source).

```
install.packages("mbX")
```

If you prefer to download the package from the GitHub repo and install it from you own computer.

Set the working directory to the one where you downloaded the package file (i.e. mbX.tar.gz) and run the following command in the R console.

```
install.packages("mbX.tar.gz", repos = NULL, type = "source")
```

Loading the package in your R session

```
library(mbX)
```

Package Workflow

1. **Data Preparation:**
    - **Taxonomic Data:** Prepare your raw 16S taxonomic data (preferably by downloading the Level 7 taxonomic level CSV file from a QIIME2 taxa barplot ".qzv" artifact, or obtain an equivalent file by exporting the appropriate artifact from QIIME2) with clear headers and sample IDs.
    - **Metadata:** Ensure your metadata file is correctly formatted (TXT, CSV, XLS, or XLSX) with a valid header and consistent sample identifiers.

2. **Data Cleaning (ezclean):**
    - Imports, validates, and standardizes your microbiome data.
    - Merges count data with metadata, ensuring consistent IDs.
    - Outputs a cleaned Excel file tailored to the selected taxonomic level (e.g., domain/kingdom, family, genus).

3. **Data Visualization (ezviz):**
    - Further processes the cleaned data by grouping, averaging, and transposing.
    - Aggregates low-abundance taxa into an "Other" category (using either a threshold or a top-taxa cutoff).
    - Generates publication-quality stacked bar plots with dynamic titles and legend labels.
    - Produces three final files in your working directory:
        1. **Cleaned Excel file** (e.g., `mbX_cleaned_families.xlsx`)
        2. **Visualization data file** (e.g., `mbX_vizualization_data_families.xlsx`)
        3. **High-resolution plot** (e.g., `mbX_viz_families.pdf`)

4. **Statistical Analysis (ezstat, Coming Soon):**

- Will offer advanced statistical workflows integrated with the same files and parameters.
- Keeps the analysis pipeline cohesive, so you don't need to switch between multiple tools for different stages of your microbiome project.

5. **Customization and Integration:**
    - The processed visualization data file offers flexibility for custom plots, scale adjustments, and color modifications.
    - Intermediate files are automatically cleaned up, leaving your working directory with only the final, essential outputs.

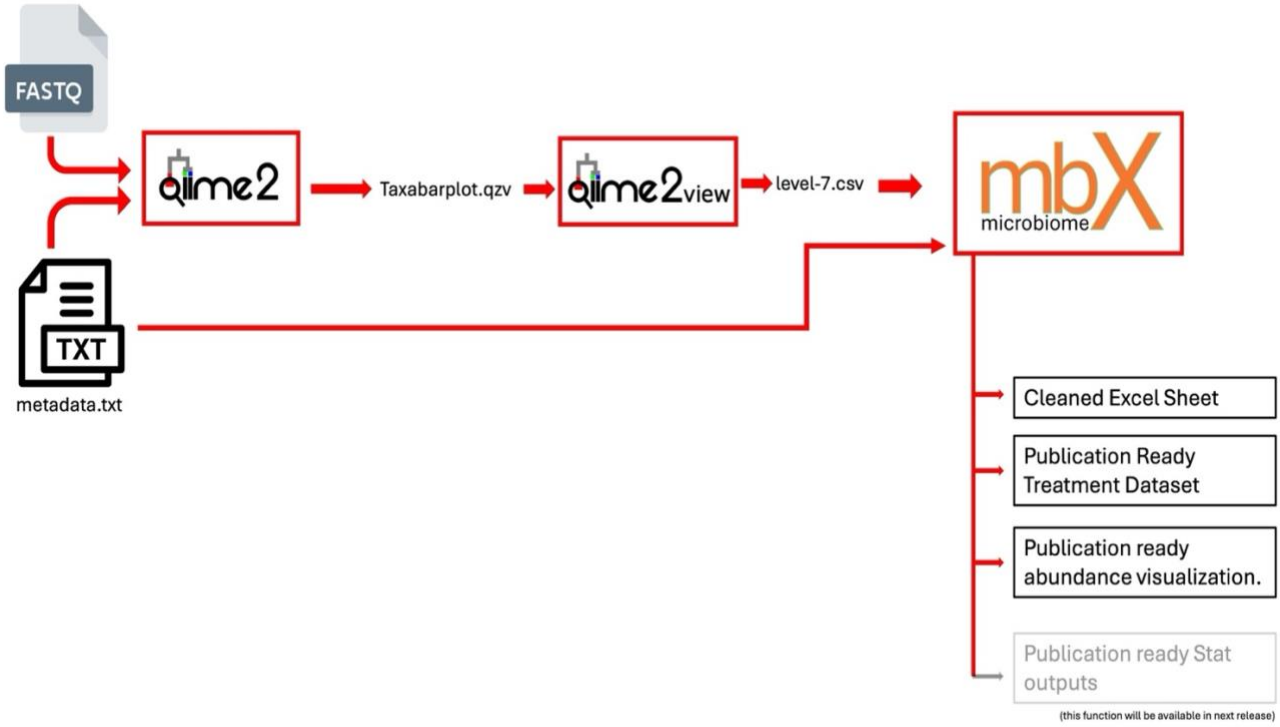

Figure 1: A high-level overview of mbX.

ezclean

ezclean

The **ezclean** function is a core component of the **mbX** package, designed to integrate seamlessly into your bioinformatics pipeline for 16S rRNA gene sequence data analysis. Whether you're working manually or automating your workflow, this function takes your raw taxonomic data and associated metadata and precisely cleans and processes the data step by step. It outputs a final Microsoft Excel file (.xlsx) that is perfectly formatted for downstream analysis. Moreover, ezclean works flawlessly with files to and from the QIIME2 pipeline, ensuring that every aspect of your data is fully automated and reproducible within your bioinformatics workflow. This makes it an essential tool for anyone looking to streamline and enhance their 16S sequence data processing in a modern bioinformatics pipeline.

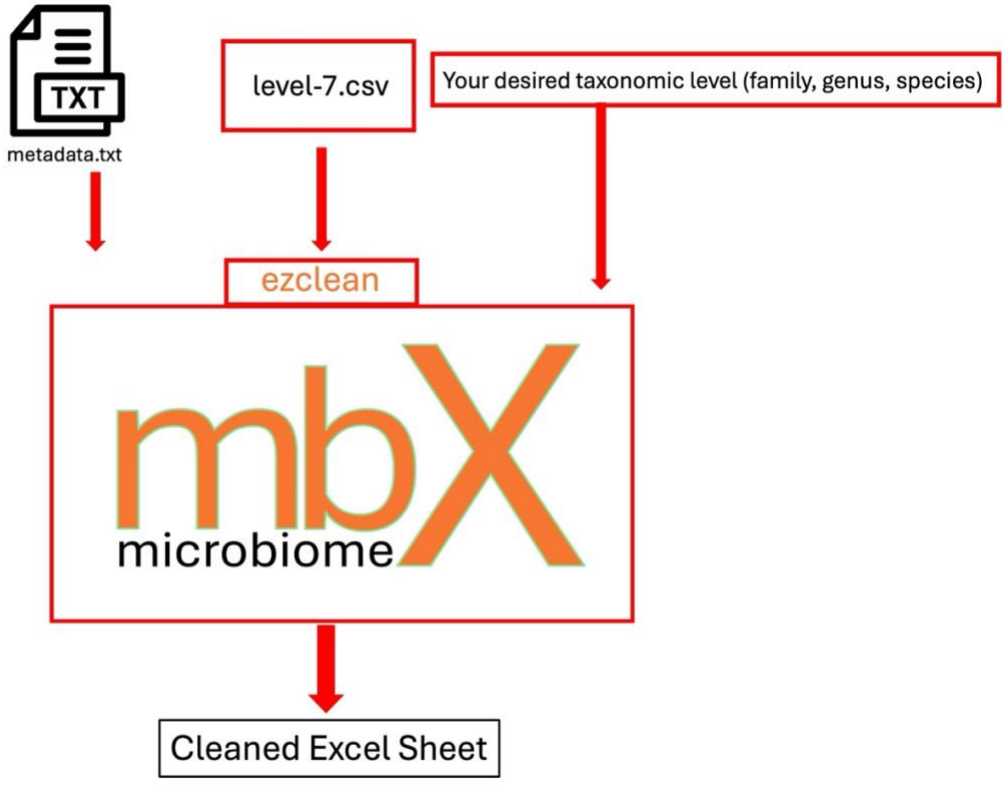

Figure 2: A high-level overview of ezclean function.

Installing the package

```
install.packages("mbX")
```

Calling the package

```
library(mbX)
```

Function calling in R (RStudio, Jupyter Notebook, VS code)

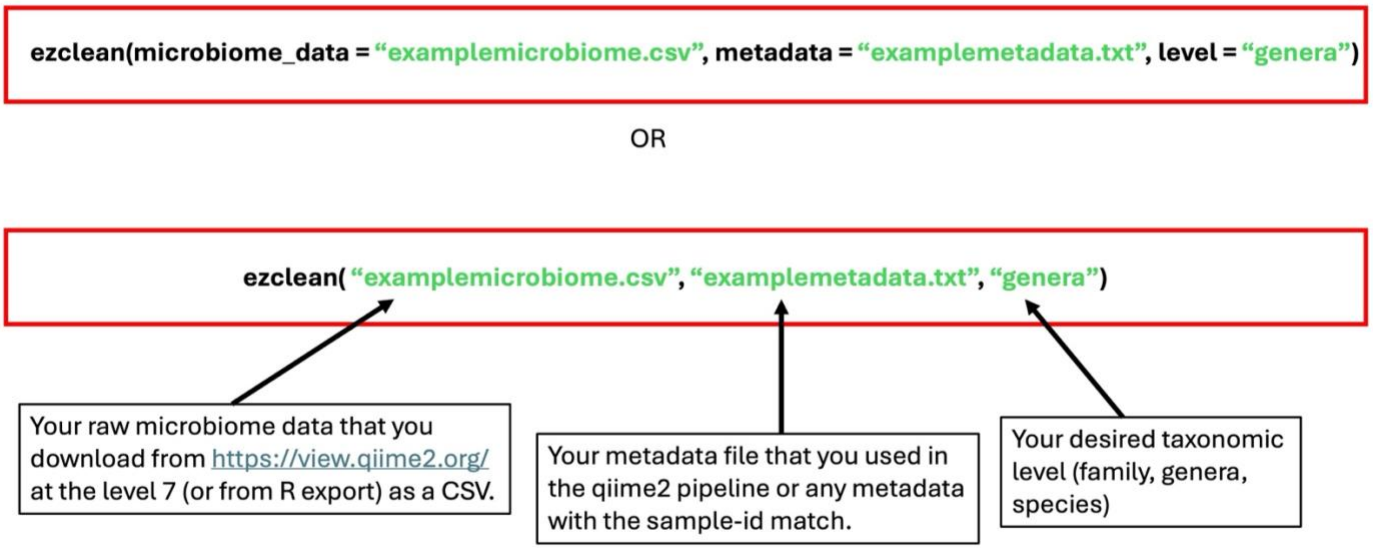

Figure 3: Example usage of the ezclean function. (must specify all taxonomic levels in last box)

Inputs and Parameters

The **ezclean** function is designed to be as straightforward as possible: it needs just **two files** and **one parameter**. First, you provide the **microbiome data** file, which contains the counts or abundances of microbial taxa across your samples. Next, you supply the **metadata** file, which holds extra details about each sample (like IDs, experimental conditions, sample types, treatments or collection dates). Finally, you specify a **taxonomic level** (for example, family or genus) so that ezclean knows how to group and clean your data. By combining these three inputs, ezclean takes care of all the heavy lifting so you can quickly move on to analyzing your 16S results.

1. **microbiome_data**

   A **CSV file** containing your raw 16S rRNA gene taxonomic data. This is the file that contains counts for each taxon across multiple samples.

Accepted Format:

- **CSV format only.**
  Although older versions of the documentation may mention TXT or Excel files, **ezclean** now strictly requires the microbiome data to be in the CSV format to ensure consistent reading and processing.

Downloading Your CSV File Directly from QIIME2 (Recommended):

- Open your QIIME2 taxa barplot ".qzv" artifact that contains all samples from your study individualized. Select "Taxonomic Level" as "Level 7", and download the CSV file.
- If you are using **QIIME2**, you can **collapse** your feature table at **level 7** (species), and **export** your collapsed table from QIIME2. Next, convert the exported table to the CSV format.
- By exporting at level 7 (or species), you ensure the file contains **all 7 taxonomic ranks**—domain (or kingdom), phylum, class, order, family, genus, and species.
- Having these multiple levels in one file gives you the **flexibility** to later select whichever taxonomic level you wish to analyze (for example, genus, or family).
- Make sure your CSV file includes clear headers (column names) and **sample IDs** that match those in your metadata file so that `ezclean` can properly merge the two datasets.

Why CSV?

- A CSV file is a universally recognized, simple text format that ensures minimal formatting conflicts, and is easier to get from the file from https://view.qiime2.org/ or R export.
- It avoids issues sometimes encountered with Excel files (hidden characters, formatting quirks, etc.) and allows the function to focus on reading the data without unexpected parsing errors.
- Even though it accepts csv, xls, xlsx or txt(tab delimited) **IT IS HIGHELY RECOMMENDED TO USE THE CSV FILE FORMAT**.

Typical Structure:

If your raw file is downloaded as CSV from https://view.qiime2.org/ or R export then you **don't need to worry about the data structure**, just use the CSV file as microbiome_data in the ezclean function. If you obtained your CSV data in another way, make sure to format your file (if necessary) to ensure it matches the same structure of CSV files that are directly downloaded from the QIIME2 taxa barplot ".qzv" artifact.

By following these guidelines, you'll set up your `microbiome_data` in a way that maximizes compatibility with `ezclean` and streamlines the downstream bioinformatics pipeline.

2. metadata

A file containing **metadata** (e.g., sample descriptions, experimental conditions, collection dates) corresponding to your microbiome samples. This additional information is crucial for grouping, comparing, and interpreting your microbiome data in context.

Accepted Formats:

- **TXT** (tab-delimited text) – ***RECOMMENDED***
- **CSV** (comma-separated values)
- **XLS** (Excel)
- **XLSX** (Excel)

Although multiple formats are supported, **TXT** is generally preferred because it is simple, avoids Excel-specific quirks, and is easy to read across different platforms.

Recommended File Use (QIIME2 Context):

- It's best to use the **same metadata file** that you employed in the **QIIME2** pipeline to maintain consistency.
- This ensures that your sample IDs and other details exactly match those in your microbiome data file.

Header Row Requirements:

- The metadata file must include a header row that defines the column names.

  Important: The column headers should **not contain spaces.**
  For example, if one of your metadata columns is named `"Sample Type"`, it is strongly recommended to rename it to `"SampleType"`.
  This practice helps avoid issues during data processing and ensures that column names are correctly recognized by the function.

Mandatory First Column Header:
The **first column header** in your metadata file must match **one** of the following terms (this is same as the compliance with the metadata format for Qiime2):

Allowed First Column Headers
id
sampleid
sample id
sample-id
featureid
feature id
feature-id

If the first column header does not match one of these exactly the `ezclean` function will stop and prompt you to correct your file.

Why the First Column Matters:

- The **first column** serves as the key that links each row of your metadata to the corresponding columns in your microbiome data file.
- Ensuring this column header matches one of the accepted terms guarantees that `ezclean` can correctly identify and merge the metadata with the microbiome data.

By following these guidelines for your metadata file—using a tab-delimited TXT format (if possible), keeping headers simple (no spaces), and ensuring the first column header is correct—you'll make the cleaning and merging process smoother, particularly when working with QIIME2-derived data.

3. level

This parameter tells the function which taxonomic level you want to analyze your data at. In microbiome analysis, taxonomic levels refer to the hierarchy of classification such as domain/kingdom, phylum, class, order, family, genus, or species. By specifying the level, you decide how the data will be grouped and summarized.

Accepted Values:
You can specify the level using either the full word or a short abbreviation. Here are the accepted inputs:

Taxonomic Level	Accepted Inputs (ideally they should all be singular: Genus)
Domain/Kingdom	"domain", "Domain", "DOMAIN", "D", "d", "kingdom", "Kingdom", "KINGDOM", "K", "k"
Phylum	"phylum", "Phylum", "PHYLUM", "P", "p"
Class	"class", "Class", "CLASS", "C", "c"
Order	"order", "Order", "ORDER", "O", "o"
Family	"family", "Family", "FAMILY", "F", "f"
Genera/Genus	"genera", "Genera", "GENERA", "G", "g"
Species	"species", "Species", "SPECIES", "S", "s"

Final Output Files

After processing, the cleaned microbiome data is saved as an Excel file. The file name is automatically determined based on the taxonomic level you selected. The table below summarizes the mapping between the taxonomic level and the final output file name:

Taxonomic Level	Final Output File Name
Domain/Kingdom	mbX_cleaned_domains_or_kingdom.xlsx
Phylum	mbX_cleaned_phylum.xlsx (ideally: “phyla”
Class	mbX_cleaned_classes.xlsx
Order	mbX_cleaned_orders.xlsx
Family	mbX_cleaned_families.xlsx
Genera/Genus	mbX_cleaned_genera.xlsx
Species	mbX_cleaned_species.xlsx

Additionally, any intermediate files created during the process are automatically deleted to keep your working directory clean.

ezviz

ezviz

The ezviz function is one of the functions from mbX package, designed to seamlessly integrate into your bioinformatics pipeline for 16S rRNA gene sequence data analysis. Beyond simply cleaning your data, ezviz validates, processes, aggregates, and generates visualizations of your microbiome data alongside its corresponding metadata. It carefully transforms raw inputs into a series of well-structured intermediate outputs, culminating in publication-quality visualizations that highlights microbial relative abundance. Whether you're manually working through your analysis or fully automating your workflow, ezviz ensures every step—from data cleaning to dynamic visualization and aggregation of low-abundance taxa—is both reproducible and optimized for modern bioinformatics (NGS datasets).

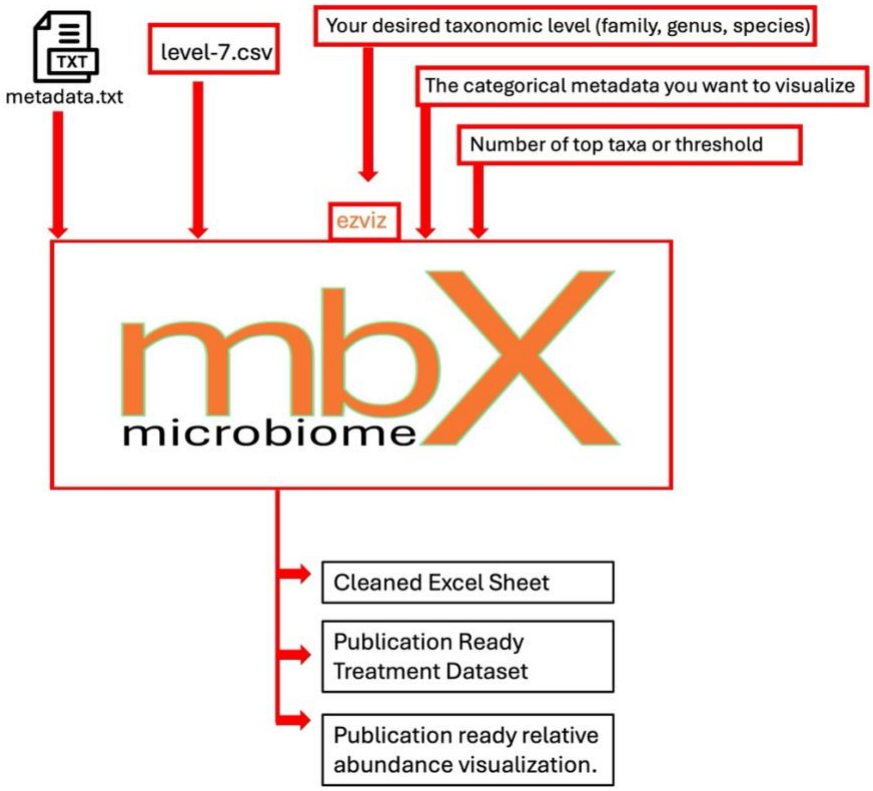

Figure 4: A high-level overview of ezviz function.

Installing the package

```
install.packages("mbX")
```

Calling the package

```
library(mbX)
```

Function calling in R (RStudio, Jupyter Notebook, VS code)

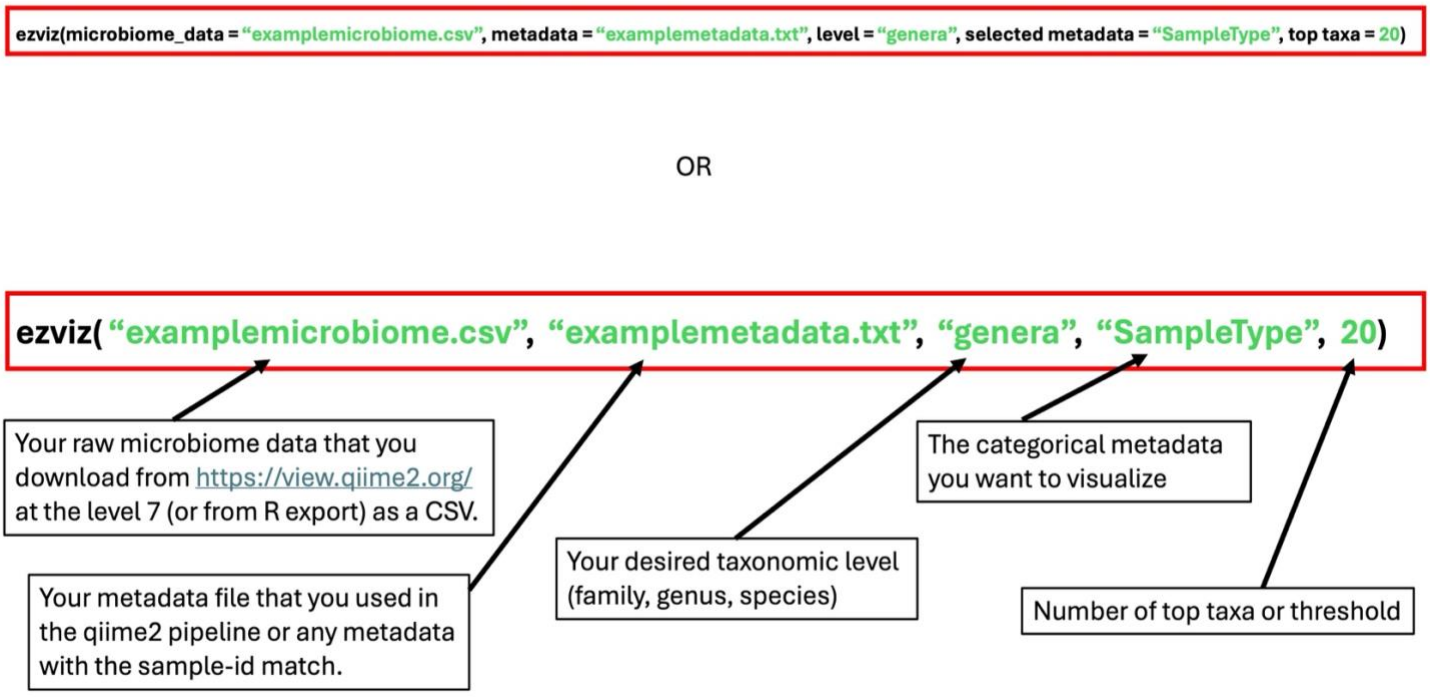

Figure 5: Example usage of the ezviz function.

Inputs and Parameters

The ezviz function follows a straightforward approach: it needs two primary files and a handful of parameters to produce a clear, publication-ready visualization. First, you supply the microbiome data file, which contains the abundance or counts of microbial taxa across

your samples. Then, you provide the metadata file, which holds additional information about each sample—like IDs, experimental conditions, sample types, treatments, or collection dates. You also specify the taxonomic level (e.g., family or genera etc.) and the particular metadata column you want to group by. Finally, you choose between a threshold or a top_taxa value to determine how less-abundant taxa are combined into an "Other" category. With these inputs, ezviz manages all the complex steps—from cleaning and aggregation to final plotting—so you can swiftly move on to analyzing and interpreting your 16S results.

1. **microbiome_data**

   A **CSV file** containing your raw 16S rRNA taxonomic data. This file typically includes counts or abundances for each taxon across multiple samples.

Accepted Format:

- **CSV only recommended.**
  Although older versions of the documentation may mention TXT or Excel files, **ezviz** now strictly requires the microbiome data to be in CSV format to ensure consistent reading and processing.

Getting Your CSV from QIIME2 (Recommended):

- Open your QIIME2 taxa barplot ".qzv" artifact that contains all samples from your study individualized. Select "Taxonomic Level" as "Level 7", and download the CSV file.
- If you are using **QIIME2**, you can **collapse** your feature table at **level 7** (species), and **export** your collapsed table from QIIME2. Next, convert the exported table to the CSV format.
- By exporting at level 7 (or species), you ensure the file contains **all 7 taxonomic ranks**—domain (or kingdom), phylum, class, order, family, genus, and species.
- Having these multiple levels in one file gives you the **flexibility** to later select whichever taxonomic level you wish to analyze (for example, genus, or family).
- Make sure your CSV file includes clear headers (column names) and **sample IDs** that match those in your metadata file so that `ezclean` can properly merge the two datasets.
- Having these multiple levels in one file gives you the **flexibility** to later select whichever taxonomic level you wish to analyze (for example, genus vs. family).
- Make sure your CSV file includes clear headers (column names) and **sample IDs** that match those in your metadata file so that `ezviz` can properly merge the two datasets.

Typical Structure:

If your raw file is downloaded as CSV from https://view.qiime2.org/ or R export then you **don't need to worry about the data structure**, just use the CSV file as microbiome_data in the ezclean function. If you obtained your CSV data in another way, make sure to format your file (if necessary) to ensure it matches the same structure of CSV files that are directly downloaded from the QIIME2 taxa barplot ".qzv" artifact.

By following these guidelines, you'll set up your `microbiome_data` in a way that maximizes compatibility with `ezviz` and streamlines the downstream bioinformatics pipeline.

2. **metadata**

A file containing **metadata** (e.g., sample descriptions, experimental conditions, collection dates) corresponding to your microbiome samples. This additional information is crucial for grouping, comparing, and interpreting your microbiome data in context.

Accepted Formats:

- **TXT** (tab-delimited text) – ***RECOMMENDED***
- **CSV** (comma-separated values)
- **XLS** (Excel)
- **XLSX** (Excel)

Although multiple formats are supported, **TXT** is generally preferred because it is simple, avoids Excel-specific quirks, and is easy to read across different platforms.

Recommended File Use (QIIME2 Context):

- It's best to use the **same metadata file** that you employed in the **QIIME2** pipeline to maintain consistency.
- This ensures that your sample IDs and other details exactly match those in your microbiome data file.

Header Row Requirements:

- The metadata file must include a header row that defines the column names.

- **Important:** The column headers should **not contain spaces.**
  For example, if one of your metadata columns is named `"Sample Type"`, it is strongly recommended to rename it to `"SampleType"`.
  This practice helps avoid issues during data processing and ensures that column names are correctly recognized by the function.

Mandatory First Column Header:
The **first column header** in your metadata file must match **one** of the following terms (this is same as the compliance with the metadata format for Qiime2):

Allowed First Column Headers
id
sampleid
sample id
sample-id
featureid
feature id
feature-id

If the first column header does not match one of these exactly the `ezviz` function will stop and prompt you to correct your file.

Why the First Column Matters:

- The **first column** serves as the key that links each row of your metadata to the corresponding columns in your microbiome data file.
- Ensuring this column header matches one of the accepted terms guarantees that `ezviz` can correctly identify and merge the metadata with the microbiome data.

By following these guidelines for your metadata file—using a tab-delimited TXT format (if possible), keeping headers simple (no spaces), and ensuring the first column header is correct—you'll make the cleaning and merging process smoother, particularly when working with QIIME2-derived data.

3. level

This parameter tells the function which taxonomic level you want to analyze your data at. In microbiome analysis, taxonomic levels refer to the hierarchy of classification such as domain/kingdom, phylum, class, order, family, genus, or species. By specifying the level, you decide how the data will be grouped and summarized.

Accepted Values:
You can specify the level using either the full word or a short abbreviation. Here are the accepted inputs:

Taxonomic Level	Accepted Inputs
Domain/Kingdom	"domain", "Domain", "DOMAIN", "D", "d", "kingdom", "Kingdom", "KINGDOM", "K", "k"
Phylum	"phylum", "Phylum", "PHYLUM", "P", "p"
Class	"class", "Class", "CLASS", "C", "c"
Order	"order", "Order", "ORDER", "O", "o"
Family	"family", "Family", "FAMILY", "F", "f"
Genera/Genus	"genera", "Genera", "GENERA", "G", "g"
Species	"species", "Species", "SPECIES", "S", "s"

4. selected_metadata

This parameter tells **ezviz** which column in your metadata file holds the categorical grouping variable for your samples. During processing, **ezviz** groups (and later averages) your microbiome data based on the unique values in this column, making it the core driver for how your samples are categorized in the final visualization.

For example, if you set `selected_metadata = "SampleType"`, the function will create stacked bars for each unique type of sample (e.g., “Soil”, “Gut”, “Water”) and compute the average microbial abundances within each category. This lets you compare relative abundances across these distinct groups in a single, publication-ready plot.

Because grouping is done by categories, the column you choose **must be a factor or categorical column**. If the column does not exist in the metadata or is numeric, **ezviz** will generate an error. Typical use cases include grouping by treatment, time point, collection site, or subject ID—basically, any attribute you want to compare in your final plot.

5. top_taxa

It is a numeric parameter.

This parameter specifies how many of the most abundant taxa you want to retain in your final visualization. When you set `top_taxa`, **ezviz** will:

1. **Calculate the mean abundance** of each taxon (across all samples).
2. **Sort the taxa in descending order** based on their mean abundances.
3. **Keep the top N taxa**, where N is the value, you assigned to `top_taxa`.
4. **Aggregate the remaining (lower-ranked) taxa** into a single row named `"Other"` (with a name tailored to the taxonomic level, e.g., `"Other_families"`).

Use `top_taxa` when you're primarily interested in highlighting the most abundant taxa, without worrying about a specific minimum threshold for inclusion. This approach helps keep your plots clear and readable by consolidating less prevalent taxa into a single category.

Important: If you also specify a `threshold`, **ezviz** will produce a message telling you to use only one because only one of these parameters can be used at a time.

6. threshold

It is a numeric parameter.

If you prefer to decide how low-abundance taxa are handled based on a minimum average abundance, use `threshold`. **ezviz** will:

1. **Compute the mean abundance** for each taxon.
2. **Compare each taxon's mean** to the specified threshold value.
3. **Retain taxa** that meet or exceed the threshold.
4. **Combine all taxa below that threshold** into a single `"Other"` row (again, named according to the taxonomic level, like `"Other_genera"` or `"Other_species"`).

Choose `threshold` if your goal is to eliminate background "noise" by filtering out taxa that do not reach a certain abundance. This lets you visualize only those taxa you deem biologically relevant.

- **Important:** If you also provide `top_taxa`, **ezviz** will produce a message telling you to use only one. Only one of these parameters can be active at a time.

Key Notes for Both Parameters (`top_taxa` and `threshold`)

- **Mutual Exclusivity:** You cannot use both `top_taxa` and `threshold` together. If both are supplied, the function will produce a message telling you to use only one.
- **Optional Usage:** If neither `top_taxa` nor `threshold` is provided, **ezviz** does **not** aggregate any taxa into an "Other" category. Instead, it visualizes all taxa as they appear after the cleaning and grouping steps. So, using one of the parameters is recommended.
- **Naming Convention:** The name of the "Other" row depends on your chosen taxonomic level (for instance, `"Other_families"` at the family level, `"Other_species"` at the species level, etc.).
- **Effect on Plot Clarity:** Applying `top_taxa` or `threshold` is highly recommended for larger datasets to keep your stacked bar plots more readable and interpretable.

Final Outputs

The `ezviz` function does not return an R object directly; instead, it produces three key files in your working directory. These outputs represent distinct stages of the workflow—from initial cleaning to final visualization—and are described below:

1. **Cleaned excel file (.xlsx)**

After processing, the cleaned microbiome data is saved as an Excel file. The file name is automatically determined based on the taxonomic level you selected. The table below summarizes the mapping between the taxonomic level and the final output file name:

Taxonomic Level	Final Output File Name
Domain/Kingdom	mbX_cleaned_domains_or_kingdom.xlsx
Phylum	mbX_cleaned_phylum.xlsx
Class	mbX_cleaned_classes.xlsx
Order	mbX_cleaned_orders.xlsx
Family	mbX_cleaned_families.xlsx
Genera/Genus	mbX_cleaned_genera.xlsx
Species	mbX_cleaned_species.xlsx

2. Visualization Data File (.xlsx)

This Excel file contains the final processed and aggregated dataset that is used for visualization. It results from further cleaning steps such as grouping by the selected metadata, averaging numeric values, transposing data, and aggregating low-abundance taxa into an "Other" category.

The file includes rows representing individual taxa (or an aggregated "Other" category, with a name dynamically generated based on the taxonomic level) and columns corresponding to samples or grouped categories.

Users have full flexibility with this data. You can use it to create alternative visualizations, adjust scales, modify legend colors, or apply custom themes using your preferred plotting tools. This file serves as a valuable resource for any further analysis or visualization needs, allowing you to tailor the display to best suit your presentation or research requirements.

The filename is generated based on the taxonomic level (e.g., `"mbX_vizualization_data_families.xlsx"` for family-level data).

Taxonomic Level	Visualization Data File Name
Domain/Kingdom	mbX_vizualization_data_domains_or_kingdom.xlsx
Phylum	mbX_vizualization_data_phyla.xlsx
Class	mbX_vizualization_data_classes.xlsx
Order	mbX_vizualization_data_orders.xlsx
Family	mbX_vizualization_data_families.xlsx
Genera/Genus	mbX_vizualization_data_genera.xlsx
Species	mbX_vizualization_data_species.xlsx

3. Visualization plot file (.pdf)

A high-resolution, publication-quality stacked bar plot is generated to visually represent the relative abundance of microbial taxa across the defined groups.

The plot includes dynamic titles, axis labels, and legends (e.g., “Relative Abundance of Microbial Families” and “Microorganism Families”) that are tailored based on your chosen taxonomic level and metadata grouping. A custom interleaved color palette is used to distinctly represent each taxon, and the dimensions of the plot are automatically adjusted based on the dataset's size.

The final plot is saved as a PDF file (ensuring optimal quality for publication) with a dynamically generated name—such as `"mbX_viz_families.pdf"` for family-level analysis.

Taxonomic Level	Visualization Plot File Name
Domain/Kingdom	mbX_viz_domains_or_kingdom.pdf
Phylum	mbX_viz_phylum.pdf
Class	mbX_viz_classes.pdf
Order	mbX_viz_orders.pdf
Family	mbX_viz_families.pdf
Genera/Genus	mbX_viz_genera.pdf
Species	mbX_viz_species.pdf

This plot serves as the final visual summary of your data analysis and can be directly used in presentations, publications, or further reporting.

Example Outputs:
After running `ezviz`, you might see the following files in your working directory:

- **Cleaned File:** `mbX_cleaned_families.xlsx`
  (The standardized and cleaned data produced by `ezclean`.)
- **Visualization Data File:** `mbX_vizualization_data_families.xlsx`
  (The final aggregated dataset ready for visualization.)
- **Visualization Plot File:** `mbX_viz_families.pdf`
  (A high-resolution, publication-ready stacked bar plot.)