Package ‘ClassifyITS’

May 7, 2026

Title Fungal Assignment Pipeline

Version 1.0.2

Description Fungi are ubiquitous in Earth's wonderfully diverse ecosystems. The 'ClassifyITS' package aids in the taxonomic classification of environmental internal transcribed spacer (ITS) short-read barcoding data. Unlike previous methods, it employs taxon-specific e-value and percent identity cutoffs at each taxonomic rank from kingdom to species. The package takes a conservative approach and outputs both graphics and user-friendly files to help users manually inspect fungal operational taxonomic units (OTUs) that fail classification at relevant levels (e.g., Phylum). 'ClassifyITS' is based on taxonomic cutoff criteria from ``The Global Soil Mycobiome consortium dataset for boosting fungal diversity research'' (Fungal Diversity, Tedersoo, 2021, <doi:10.1007/s13225-021-00493-7>) and ``Best practices in metabarcoding of fungi: From experimental design to results'' (Molecular Ecology, Tedersoo, 2022, <doi:10.1111/mec.16460>).

License GPL-3

Encoding UTF-8

Imports ggplot2, gridExtra, grid, reshape2, data.table, seqinr

Suggests formatR, knitr, rmarkdown

RoxygenNote 7.3.3

VignetteBuilder knitr, rmarkdown, formatR

NeedsCompilation no

Author Quinn Moon [aut, cre]

Maintainer Quinn Moon <`qmoon@umich.edu`>

Repository CRAN

Date/Publication 2026-04-23 15:40:02 UTC

Contents

`best_hit_taxonomy_assignment`

Hierarchical best-hit taxonomy assignment with per-rank fallback rule

---

Description

Pass ONLY those OTUs that haven't been assigned already! For each rank, if the best e-value hit is undefined and the second-best hit is defined and at least 60

Usage

```
best_hit_taxonomy_assignment(
  blast_qc,
  cutoffs_long,
  genus_cutoff_mode = c("prefer_evalue", "prefer_pident", "both")
)
```

Arguments

`blast_qc`	A data.frame of BLAST results for query sequences. Must include qseqid, evalue, pident, length, and taxonomy columns: kingdom/phylum/class/order/family/genus/species.
`cutoffs_long`	Long-form cutoffs (parse_taxonomy_cutoffs()$long).
`genus_cutoff_mode`	One of: "prefer_evalue", "prefer_pident", "both".

Details

Defaults are taken from the cutoffs table itself (Fungi baseline rules), not from a separate defaults list.

Value

A data.frame containing hierarchical taxonomy assignment for each query sequence.

`check_N` *Check proportion of N bases in each sequence.*

Description

Calculates the proportion of "N" bases (ambiguous bases) in each sequence and flags if above the given threshold.

Usage

```
check_N(rep_seqs, cutoff = 1)
```

Arguments

`rep_seqs`	Character vector, list (e.g., from seqinr::read.fasta(as.string=TRUE)), or (optionally) a DNAStringSet.
`cutoff`	Numeric, percent threshold (default 1).

Value

Data frame with columns: qseqid, N_percent, N_flag.

Examples

```
seqs <- c(seq1 = "ATGCNNNN", seq2 = "NNNNATGC")
check_N(seqs)
check_N(seqs, cutoff = 10)
```

`consensus_taxonomy_assignment` *Per-rank consensus filter for taxonomy assignment*

Description

Only confirms or demotes, never promotes Unclassified. FINAL hierarchy check: if any rank is Unclassified, all lower ranks are forced to Unclassified.

Usage

```
consensus_taxonomy_assignment(final_table, blast_qc)
```

Arguments

`final_table`	Data frame of taxonomic assignments.
`blast_qc`	Data frame of filtered BLAST hits for each OTU.

Value

Data frame of consensus assignments (same structure as input).

`easy_assignments` *Easy taxonomy assignment for OTUs using BLAST QC output & phylum-specific thresholds.*

Description

Easy taxonomy assignment for OTUs using BLAST QC output & phylum-specific thresholds.

Usage

```
easy_assignments(blast_filtered, cutoffs_file = NULL, default_cutoff = 98)
```

Arguments

`blast_filtered`	QC-filtered BLAST dataframe (with parsed taxonomy columns!)
`cutoffs_file`	Path to taxonomy cutoffs CSV file. If not supplied or invalid, attempts to locate the default file in the package.
`default_cutoff`	Default percent identity cutoff (kept for API compatibility)

Value

List with assigned_otus_df and remaining_otus_df

`ensure_cols` *Ensure data frame has all required columns (as character)*

Description

Ensure data frame has all required columns (as character)

Usage

```
ensure_cols(df, all_cols)
```

Arguments

`df`	Data frame to fix
`all_cols`	Vector of required columns

Value

Fixed data frame (in correct order, with all columns present)

`ITS_assignment` *Complete Fungal Assignment Pipeline*

Description

Runs all steps: QC, filtering, plotting, assignments; optionally writes outputs.

Usage

```
ITS_assignment(
  blast_file,
  rep_fasta,
  cutoffs_file = NULL,
  cutoff_fraction = 0.6,
  n_cutoff = 1,
  outdir = NULL,
  verbose = FALSE
)
```

Arguments

`blast_file`	Path to BLAST results TSV file
`rep_fasta`	Path to representative sequences FASTA file
`cutoffs_file`	Path to taxonomy cutoffs CSV file (optional; defaults to package example if omitted)
`cutoff_fraction`	Numeric, fraction of median rep-seq length for BLAST filtering (default: 0.6)
`n_cutoff`	Numeric, N base percentage cutoff (default: 1)
`outdir`	Output directory for results. If NULL (default), nothing is written.
`verbose`	Logical; if TRUE emit progress messages. Default FALSE.

Value

Named list of results and (if written) output file paths

`load_and_check` *Load and check BLAST results and rep-seq FASTA*

Description

Load and check BLAST results and rep-seq FASTA

Usage

```
load_and_check(blast_file, rep_fasta, taxonomy_col = "stitle", verbose = FALSE)
```

Arguments

`blast_file`	Path to BLAST results TSV file.
`rep_fasta`	Path to representative sequences FASTA file.
`taxonomy_col`	The column in BLAST file containing taxonomy strings (default "stitle").
`verbose`	Logical; if TRUE, emit progress messages. Default FALSE.

Value

List with BLAST dataframe (kingdom-filtered) and rep_seqs as a named list of DNA strings.

`parse_taxonomy_cutoffs` *Parse taxonomy cutoffs file*

Description

Reads and processes a taxonomy cutoffs CSV for assignment thresholds at various ranks.

Usage

```
parse_taxonomy_cutoffs(cutoffs_file = NULL)
```

Arguments

`cutoffs_file`	Path to a taxonomy cutoffs CSV file. If not supplied or invalid, attempts to locate the default file in the package.

Value

A list with two elements: `long`, a data frame of parsed cutoffs, and `ranks`, the vector of taxonomic ranks.

`plot_alignment_hist` *Create and return alignment length histogram (ggplot object)*

Description

Create and return alignment length histogram (ggplot object)

Usage

```
plot_alignment_hist(blast, rep_seqs, cutoff_fraction = 0.6)
```

Arguments

`blast`	BLAST data frame.
`rep_seqs`	Named list/character vector of DNA strings (from seqinr::read.fasta(as.string = TRUE)).
`cutoff_fraction`	Numeric; fraction of median alignment length for cutoff line. Default 0.6.

Value

A ggplot object.

`safe_rbind_list` *Safely rbinds list of data frames, ensuring columns match*

Description

Safely rbinds list of data frames, ensuring columns match

Usage

```
safe_rbind_list(dfs, all_cols = NULL)
```

Arguments

`dfs`	List of data frames
`all_cols`	Vector of required columns

Value

Combined data frame

`save_taxonomy_graphics` *Save taxonomy summary charts and tables to multi-page PDF*

Description

Save taxonomy summary charts and tables to multi-page PDF

Usage

```
save_taxonomy_graphics(
  all_results,
  hist_plot,
  pdf_file = NULL,
  caption_texts = NULL,
  rank_names = c("Phylum", "Class", "Order", "Family", "Genus", "Species"),
  verbose = FALSE
)
```

Arguments

`all_results`	Combined assignments table from write_initial_assignments
`hist_plot`	ggplot2 object for histogram
`pdf_file`	Output path for multi-page PDF. If NULL (default), no file is written.
`caption_texts`	Vector of captions for PDF pages (optional)
`rank_names`	Vector of rank names (default: c("Phylum",...))
`verbose`	Logical; if TRUE, emit a message when a PDF is written. Default FALSE.

Value

List with plots/tables; includes pdf_file when written.

`trim_alignments` *Trim BLAST alignments by minimum length*

Description

Trim BLAST alignments by minimum length

Usage

```
trim_alignments(blast, rep_seqs, fraction = 0.6)
```

Arguments

`blast`	BLAST data frame.
`rep_seqs`	Named list/character vector of DNA strings (from seqinr::read.fasta(as.string = TRUE)).
`fraction`	Numeric; fraction of the median rep-seq length used as the cutoff. Default 0.6.

Value

Filtered BLAST data frame.

`write_initial_assignments` *Create and write the initial assignments table including drops at all steps*

Description

Create and write the initial assignments table including drops at all steps

Usage

```
write_initial_assignments(
  easy_df,
  consensus_df,
  rep_seqs,
  blast,
  blast_filtered,
  file = NULL,
  verbose = FALSE
)
```

Arguments

`easy_df`	Data frame of easy-assigned OTUs
`consensus_df`	Data frame of consensus-assigned OTUs (hard ones)
`rep_seqs`	DNAStringSet or named character vector of rep seqs
`blast`	Data frame of all BLAST results
`blast_filtered`	Data frame of filtered BLAST results
`file`	Path for output CSV. If NULL (default), no file is written.
`verbose`	Logical; if TRUE emit a message when a file is written. Default FALSE.

Value

Data frame of assignments (written if file is not NULL)

Index