# Package: malaviR (via r-universe)

June 8, 2026

**Type** Package

**Title** An R interface to MalAvi

**Version** 1.0.0

**Maintainer** Vincenzo A. Ellis <vaellis@udel.edu>

**Description** Functions for working with data from the MalAvi database of avian haemosporidian parasites.

**License** GPL-2

**Encoding** UTF-8

**LazyData** true

**Depends** R (>= 3.5.0)

**Imports** ape, dplyr, magrittr, utils

**Suggests** Biostrings, clootl, DECIPHER (>= 3.0), readxl, testthat

**RoxygenNote** 7.3.1

**URL** https://github.com/vincenzoaellis/malaviR

**BugReports** https://github.com/vincenzoaellis/malaviR/issues

**Repository** https://vincenzoaellis.r-universe.dev

**Date/Publication** 2026-06-08 16:28:36 UTC

**RemoteUrl** https://github.com/vincenzoaellis/malaviR

**RemoteRef** HEAD

**RemoteSha** b9d3de84d102a6ba5c58cf7f1fb3986ad5173051

## Contents

---

`blast_malavi` *BLAST-like search of a sequence against MalAvi*

---

## Description

Finds the MalAvi lineages most similar to a query DNA sequence against the database bundled in the package. This uses **DECIPHER**: the bundled, pre-built inverted index is searched with `DECIPHER::SearchIndex` and the top hits are aligned to the query with `DECIPHER::AlignPairs`.

## Usage

```
blast_malavi(sequence, top_n = 5, version = "latest")
```

## Arguments

| | |
|---|---|
| `sequence` | A DNA sequence as a single character string. Whitespace and gap (`-`) characters are removed; the sequence may be upper or lower case. |
| `top_n` | Number of top hits to return (default 5). |
| `version` | MalAvi release to search, as a date string (e.g. `"2026-03-23"`) or `"latest"` (default). |

## Details

**DECIPHER** (>= 3.0) and **Biostrings** are required and must be installed from Bioconductor: `BiocManager::install(c("DE "Biostrings"))`. DECIPHER >= 3.0 needs R >= 4.4.

## Value

A `data.frame` of hits, best first, with columns `Lineage`, `ProportionMatch`, `PercentMatch`, `AlignmentLength`, `Matches`, `Mismatches`, `Score`, `QueryGapLength`, `ReferenceLineageLength`, and `ReferenceFullLength`. `ReferenceLineageLength` is the position in the reference lineage where the alignment ends (as reported by the original MalAvi BLAST app), whereas `ReferenceFullLength` is the full length of the reference lineage sequence; the two differ when the query aligns to only part of a reference. If no hits are found, a one-row data frame of `NA`s is returned with a warning.

## See Also

`extract_alignment`

## Examples

```
## Not run:
## requires DECIPHER (>= 3.0) and Biostrings
seq <- paste(as.character(extract_alignment()[1, ]), collapse = "")
blast_malavi(seq, top_n = 5)

## End(Not run)
```

---

`clean_alignment` *Identify and collapse repeated haplotypes in a MalAvi alignment*

---

## Description

Shorter MalAvi lineages (i.e., < 479 bp) sometimes match perfectly to longer sequences that have different lineage names ("synonymies"), and it has been pointed out in the literature that this inflates estimates of parasite diversity (Tamayo-Quintero et al. 2025). This function finds groups of lineages that share a haplotype, returns a table of those synonymies, and produces a de-duplicated alignment that keeps one lineage per group.

## Usage

```
clean_alignment(
  alignment,
  method = c("overlap", "strict"),
  select = c("complete", "random"),
  keep = NULL
)
```

## Arguments

| | |
|---|---|
| `alignment` | A `DNAbin` alignment (e.g. from `extract_alignment`). |
| `method` | How to define a repeated haplotype: `"overlap"` (default) or `"strict"` (see Details). |
| `select` | How to pick the lineage kept from each synonymy group when it is not named in `keep`: `"complete"` (default) keeps the most complete sequence (ties broken alphabetically); `"random"` keeps one at random (set a seed first for reproducibility). |
| `keep` | Optional character vector of lineage names to keep. For each synonymy group containing one of these names, that name is kept; an error is raised if a single group contains more than one supplied name. |

## Details

By default this function is deterministic: the most complete (i.e., longest) sequence in each group is kept (ties broken alphabetically). Set `select = "random"` for the quick random selection of the earlier `malaviR` version, which keeps one lineage per group at random (call `set.seed` first for

reproducibility). In either case, supply `keep` to override the choice for specific groups (i.e., if you want to choose particular lineages to represent a haplotype group); any group without a supplied choice falls back to the `select` rule.

Two definitions of "same haplotype" are available via `method`:

`"overlap"` **(default)** collapses a partial sequence into any strictly more complete sequence that is identical to it over the partial's informative (non-gap/non-N) positions, in addition to collapsing fully identical sequences. This catches the partial-sequence synonymies highlighted by Tamayo-Quintero et al. (2025), but is slower on large alignments.

`"strict"` collapses only sequences that are identical across the whole alignment, including gaps – the behavior of the original function from the earlier `malaviR` version.

The `informative_length` column (count of A/C/G/T bases) helps flag the short, partial sequences at the heart of the problem.

## Value

A list with elements:

`synonymies` a `data.frame`, one row per lineage in a repeated-haplotype group, with columns `haplotype` (group id), `lineage`, `informative_length`, and `status` (`"kept"` or `"dropped"`).

`kept` character vector of lineages kept.

`dropped` character vector of lineages dropped.

`alignment_clean` the `DNAbin` alignment with dropped lineages removed.

## References

Tamayo-Quintero J, Martinez-de la Puente J, Matta NE, Pacheco MA, Rivera-Gutierrez HF (2025). Imprudent use of MalAvi names biases the estimation of parasite diversity of avian haemosporidians. PLoS Pathogens 21(2): e1012911. doi:10.1371/journal.ppat.1012911

## See Also

`synonymy_report`, `extract_alignment`

## Examples

```
aln <- extract_alignment()
res <- clean_alignment(aln)
head(res$synonymies)

## quick random pick (reproducible with a seed)
set.seed(1)
res_rand <- clean_alignment(aln, select = "random")
```

clean_names *Clean MalAvi lineage names to match the tables*

## Description

MalAvi alignment tip labels carry a parasite-genus prefix (e.g. "H_COLL2"), and often a trailing morphological-species name as well (e.g. "H_COLL2_Haemoproteus_pallidus"), whereas the data tables store the lineage name alone (e.g. "COLL2"). This helper strips the prefix and any trailing morphological-species name so names from an alignment can be matched to the tables, and can optionally return the parasite genus alongside the cleaned name.

## Usage

```
clean_names(lin.names, keep.genus = FALSE)
```

## Arguments

| | |
|---|---|
| lin.names | Character vector of lineage names of the form "<genus prefix>_<lineage>", optionally followed by a morphological-species name (e.g. from rownames() of an alignment). |
| keep.genus | If FALSE (default), return just the cleaned lineage names as a character vector. If TRUE, return a data.frame with the parasite genus (P/H/L expanded to *Plasmodium*/*Haemoproteus*/*Leucocytozoon*) and the cleaned Lineage_Name. |

## Value

A character vector, or a data.frame when keep.genus = TRUE.

## Examples

```
clean_names(c("H_COLL2_Haemoproteus_pallidus", "P_GRW04_Plasmodium_relictum", "L_CIAE02"))
clean_names(c("H_COLL2_Haemoproteus_pallidus", "L_CIAE02"), keep.genus = TRUE)
```

---

clootl_taxonomy_version *Version of the clootl taxonomy bundled in the package*

## Description

match_taxonomy matches host names against a snapshot of the **clootl** (eBird/Clements) avian taxonomy that is bundled with malaviR. This returns the taxonomy year of that snapshot.

## Usage

```
clootl_taxonomy_version()
```

## Value

The bundled clootl taxonomy year (an integer).

## References

McTavish EJ, Gerbracht JA, Holder MT, Iliff MJ, Lepage D, Rasmussen PC, Redelings BD, Sanchez Reyes LL, Miller ET (2025). A complete and dynamic tree of birds. Proceedings of the National Academy of Sciences 122(18): e2409658122. doi:10.1073/pnas.2409658122

## See Also

`match_taxonomy`

## Examples

```
clootl_taxonomy_version()
```

---

# extract_alignment *Get the MalAvi sequence alignment*

---

## Description

Returns the aligned MalAvi cytochrome *b* sequences from the database bundled in the package, as a `DNAbin` object. MalAvi is no longer downloaded from the web; the alignment comes from the release shipped with `malaviR` (see `malavi_version`).

## Usage

```
extract_alignment(
  version = "latest",
  genus = c("all", "Plasmodium", "Haemoproteus", "Leucocytozoon", "other")
)
```

## Arguments

| | |
|---|---|
| `version` | MalAvi release to read, as a date string (e.g. `"2026-03-23"`) or `"latest"` (default). |
| `genus` | Parasite genus/genera to return. Either `"all"` (default, the whole alignment) or one or more of `"Plasmodium"`, `"Haemoproteus"`, `"Leucocytozoon"`, and `"other"`. |

## Details

Lineage names are prefixed by parasite genus: `P_` (*Plasmodium*), `H_` (*Haemoproteus*), `L_` (*Leucocytozoon*); any other prefix is treated as `"other"`. Use `genus` to subset the alignment to one or more genera. Note that some tip labels also carry a morphological species name appended after the lineage code (e.g. `"H_COLL2_Haemoproteus_pallidus"`).

### Value

A DNAbin alignment, optionally subset by genus.

### See Also

extract_table, clean_alignment

### Examples

```
aln <- extract_alignment()
dim(aln)
plas <- extract_alignment(genus = "Plasmodium")
```

---

## extract_table *Get a MalAvi data table*

---

### Description

Returns one of the MalAvi data tables from the database bundled in the package. MalAvi is no longer downloaded from the web; the tables come from the release shipped with malaviR (see malavi_version).

### Usage

```
extract_table(table = "Hosts and Sites Table", version = "latest")
```

### Arguments

| | |
|---|---|
| table | Name of the table to return (see Details), or "all" to return a named list of all five tables. Defaults to "Hosts and Sites Table". |
| version | MalAvi release to read, as a date string (e.g. "2026-03-23") or "latest" (default). |

### Details

The bundled release provides five tables:

"Hosts and Sites Table" host records, sites, and references (hosts_and_sites).

"Grand Lineage Summary" per-lineage summary, including the sequence (grand_lineage_summary).

"Morpho Species Summary" lineages linked to morphologically described species (morpho_species).

"Table of References" reference list (references).

"Vector Data Table" vector records (vector_data).

Either the descriptive name above or its short snake_case key may be supplied.

### Value

A data.frame, or for table = "all" a named list of data.frames.

## See Also

extract_alignment, malavi_version

## Examples

```
hosts <- extract_table("Hosts and Sites Table")
head(hosts)
```

---

`malavi_version` *MalAvi database version bundled in the package*

---

## Description

Returns the version (release date) of the MalAvi database that `malaviR` reads from. MalAvi is no longer permanently online, so the "version" is simply the date stamp of the bundled release (e.g. `"2026-03-23"`). Use `which = "all"` to list every release bundled in your installation (currently there is only one, but I may keep some archived older versions in the future).

## Usage

```
malavi_version(which = c("latest", "all"))
```

## Arguments

| | |
|---|---|
| `which` | Either `"latest"` (default) to return the most recent bundled release, or `"all"` to return all bundled releases. |

## Value

A character vector of version (date) string(s).

## See Also

extract_table, extract_alignment

## Examples

```
malavi_version()
```

`match_taxonomy` *Match host species names to the clootl (eBird) avian taxonomy*

## Description

Aligns a set of bird species names to the avian taxonomy used by the **clootl** package (the eBird/Clements taxonomy that underlies the constantly updated avian phylogeny of McTavish et al. 2025). For each name it returns the matching eBird species, the corresponding tip label in the clootl phylogeny (`ott_name`), and the order and family, together with a `match_type` describing how (or whether) it matched.

## Usage

```
match_taxonomy(species = NULL, version = "latest", family = NULL, order = NULL)
```

## Arguments

| | |
|---|---|
| `species` | Character vector of species names to match. If `NULL` (default), the unique host species in the bundled MalAvi `"Hosts and Sites Table"` are used, along with their MalAvi family and order. |
| `version` | MalAvi release to take host names from when `species` is `NULL`; a date string or `"latest"` (default). |
| `family, order` | Optional character vectors, the same length as `species`, giving each name's family and order. They are used only for the family/order-constrained epithet step (see Details) and are taken from MalAvi automatically when `species` is `NULL`. If you supply your own `species` without them, that recovery step is simply skipped. |

## Details

Names are first looked up in a maintainer-curated override key (`data-raw/manual_taxonomy.csv`) of MalAvi host names that have been hand-resolved to a current eBird species; these are flagged `match_type = "manual"`. Remaining names are matched against the eBird scientific names, and, failing that, against the IOC, BirdLife, and Howard & Moore synonyms carried by clootl (which are then resolved back to the eBird name). Many MalAvi host names are older binomials that no longer match any of those because the genus has since been split or the specific epithet re-gendered (e.g. *Anas clypeata* is now *Spatula clypeata*; *Basileuterus basilicus* is now *Myiothlypis basilica*). To recover these, a final step matches the specific epithet – allowing for Latin gender agreement – within the host's MalAvi family (or, if that family name is not used by clootl, within its order), accepting the match only when it points to a single eBird species. This resolves most genus reassignments while the family/order constraint guards against epithet collisions between unrelated birds; names whose epithet remains ambiguous are left unmatched rather than guessed. As a last step, host names still unmatched are looked up in the hand-curated species key from the original `malaviR` (which mapped many MalAvi names to corrected binomials); the corrected name is then resolved to the current eBird name and flagged `match_type = "legacy"`. These legacy matches come from a hand-curated key made years ago against the Jetz *et al.* (BirdTree) taxonomy and may reflect taxonomic decisions that are now out of date, so they are worth double-checking.

Leading/trailing whitespace is removed before matching. Some MalAvi host names are not identifiable binomials – entries ending in "sp.", hybrids written with " x ", or bare genus names – and can never match; these are flagged `match_type = "generic"` rather than forced to a species.

The clootl taxonomy is bundled with `malaviR` as a dated snapshot, so no internet connection or **clootl** installation is needed at run time. See `clootl_taxonomy_version` for the bundled taxonomy year.

## Value

A list with two data frames:

`key` one row per input species, with columns `malavi_species`, `ebird_species`, `ott_name`, `order`, `family`, and `match_type` (one of `"manual"`, `"exact"`, `"synonym:IOC"`, `"synonym:BirdLife"`, `"synonym:HowardMoore"`, `"reassigned:family"`, `"reassigned:order"`, `"legacy"`, `"generic"`, or `"none"`).

`differences` the subset of `key` that did not match an eBird name exactly (manual overrides, synonyms, reassignments, legacy matches, generics, and unmatched names) – the rows worth checking by hand.

## References

McTavish EJ, Gerbracht JA, Holder MT, Iliff MJ, Lepage D, Rasmussen PC, Redelings BD, Sanchez Reyes LL, Miller ET (2025). A complete and dynamic tree of birds. Proceedings of the National Academy of Sciences 122(18): e2409658122. doi:10.1073/pnas.2409658122

## See Also

`taxonomy` for the pre-built crosswalk of MalAvi hosts, `clootl_taxonomy_version`

## Examples

```
res <- match_taxonomy(c("Turdus merula", "Cyanistes caeruleus", "Anas sp."))
res$key
res$differences
```

---

# `sister_taxa` *Identify the sister taxa at a node in a phylogeny*

---

## Description

For an internal node, returns the tips descending from each of its two immediate descendant clades, labelled as sister clade 1 or 2. This is useful, for example, for comparing the hosts or traits of sister lineages in a parasite phylogeny (Ellis and Bensch 2018). One or several nodes may be supplied.

## Usage

```
sister_taxa(tree, node)
```

### Arguments

| | |
|---|---|
| tree | A phylogeny of class phylo (see **ape**). |
| node | An internal node number, or a vector of node numbers. For a vector, results for each node are stacked into one data frame. |

### Value

A data.frame with columns ancestral.node, sister.clade (1 or 2), and taxa (tip label).

### References

Ellis VA, Bensch S (2018). Host specificity of avian haemosporidian parasites is unrelated among sister lineages but shows phylogenetic signal across larger clades. International Journal for Parasitology 48: 897-902. doi:10.1016/j.ijpara.2018.05.005

### Examples

```
tree <- ape::read.tree(text = "((A,B),(C,(D,E)));")
sister_taxa(tree, node = 8)
```

---

## synonymy_report *Quantify MalAvi haplotype synonymies for investigation*

---

### Description

Summarizes how many lineage names share a haplotype with another name and returns the lineage names in groups so they can be examined. By default it reports on the bundled MalAvi alignment using the overlap-aware definition of a haplotype (which catches short, partial sequences identical to a longer one), but any alignment and either method may be used.

### Usage

```
synonymy_report(
  alignment = NULL,
  method = c("overlap", "strict"),
  version = "latest"
)
```

### Arguments

| | |
|---|---|
| alignment | A DNAbin alignment. If NULL (default), the bundled MalAvi alignment for version is used. |
| method | How to define a repeated haplotype: "overlap" (default) or "strict". See clean_alignment. |
| version | MalAvi release to use when alignment is NULL; a date string or "latest" (default). |

## Details

This is a reporting companion to `clean_alignment`: use this to see the size of the problem and which lineages to check, and `clean_alignment` to actually produce a de-duplicated alignment.

## Value

A list with:

`summary` a one-row `data.frame` of counts: `n_sequences`, `n_haplotypes` (distinct haplotypes), `n_synonymous_haplotypes` (haplotypes carrying >1 lineage name), `n_lineages_in_synonymies`, `n_redundant_names` (`n_sequences - n_haplotypes`, the diversity inflation), `pct_diversity_inflation`, and `n_partial_sequences`.

`by_genus` redundant-name counts split by parasite genus.

`synonymies` a `data.frame` of the synonymy groups, one row per lineage, with `haplotype`, `lineage`, `genus`, `informative_length`, `is_partial`, and `status` – the list of names to investigate.

## References

Tamayo-Quintero J, Martinez-de la Puente J, Matta NE, Pacheco MA, Rivera-Gutierrez HF (2025). Imprudent use of MalAvi names biases the estimation of parasite diversity of avian haemosporidians. PLoS Pathogens 21(2): e1012911. doi:10.1371/journal.ppat.1012911

## See Also

`clean_alignment`, `extract_alignment`

## Examples

```
rep <- synonymy_report(method = "strict")
rep$summary
head(rep$synonymies)
```

---

# `taxonomy` *MalAvi host species matched to the clootl (eBird) taxonomy*

---

## Description

A key linking the unique host species names in the bundled MalAvi "Hosts and Sites Table" to the avian taxonomy used by the **clootl** package (the eBird/Clements taxonomy underlying the McTavish et al. 2025 avian phylogeny). It is produced by `match_taxonomy`.

## Usage

```
taxonomy
```

## Format

A data frame with one row per unique MalAvi host species and the following columns:

**malavi_species** host species name as it appears in MalAvi.

**ebird_species** matched eBird scientific name, or NA.

**ott_name** matching tip label in the clootl phylogeny, or NA.

**order** taxonomic order of the matched species, or NA.

**family** taxonomic family of the matched species, or NA.

**match_type** how the name matched: "manual", "exact", "synonym:IOC", "synonym:BirdLife", "synonym:HowardMoore", "reassigned:family", "reassigned:order", "legacy", "generic", or "none".

## Details

The bundled clootl taxonomy year is reported by clootl_taxonomy_version. The match_type column records how each name matched: "manual" was hand-resolved by the package maintainer (data-raw/manual_taxonomy.csv); "exact" matched an eBird scientific name directly; "synonym:IOC", "synonym:BirdLife", and "synonym:HowardMoore" matched via the IOC, BirdLife, or Howard & Moore names that clootl carries; "reassigned:family" and "reassigned:order" matched by specific epithet (allowing for Latin gender agreement) within the host's MalAvi family or order, recovering genus reassignments; "legacy" matched via the hand-curated key from the original malaviR (an old, possibly out-of-date choice worth double-checking); "generic" are names that cannot map to a single species (e.g. ending in "sp." or hybrids); and "none" are binomials with no match in the bundled taxonomy.

## Source

MalAvi (https://wimanet-science.github.io/web/malavi/) host species matched to the clootl taxonomy (https://github.com/eliotmiller/clootl).

## References

McTavish EJ, Gerbracht JA, Holder MT, Iliff MJ, Lepage D, Rasmussen PC, Redelings BD, Sanchez Reyes LL, Miller ET (2025). A complete and dynamic tree of birds. Proceedings of the National Academy of Sciences 122(18): e2409658122. doi:10.1073/pnas.2409658122

## See Also

match_taxonomy, clootl_taxonomy_version

# Index