

# Package ‘moultdbtools’

January 11, 2022

**Title** Combined analysis of NCBI taxonomy and GBIF taxonomy

**Version** 0.0.0.9000

**Description** The NCBI is the go to resource for taxonomic studies but it only contains data on extant species whereas the GBIF also includes extinct species. ‘moultdbtools’ is useful for the combined analysis of the NCBI taxonomy and GBIF taxonomy.

**License** CC0

**Encoding** UTF-8

**Roxygen** list(markdown = TRUE)

**RoxygenNote** 7.1.2

**Suggests** testthat (>= 3.0.0)

**Config/testthat/edition** 3

**Imports** purrr,  
dplyr,  
vroom,  
ggplot2,  
rje

**Depends** R (>= 2.10)

**LazyData** true

## R topics documented:

get_lineages . . . . .	2
get_status . . . . .	2
get_validity . . . . .	3
load_population . . . . .	3
load_sample . . . . .	4
load_taxonomies . . . . .	4
plot_mdb . . . . .	5
prepare_comparable_rank_dist . . . . .	5
prepare_rank_dist . . . . .	6
term_conversion . . . . .	6
<b>Index</b>	<b>8</b>

---

get_lineages	<i>Get entries that have lineage information for both the GBIF and NCBI data</i>
--------------	--

---

**Description**

Get entries that have lineage information for both the GBIF and NCBI data

**Usage**

```
get_lineages(x)
```

**Arguments**

**x** A tibble created with `load_taxonomies()` or `load_population()` or `load_sample()`.

**Value**

A tibble with complete lineage data

**Examples**

```
get_lineages(load_sample())
```

---

get_status	<i>Filter a combined taxonomy by GBIF taxonomic status/synonym</i>
------------	--

---

**Description**

Filter a combined taxonomy by GBIF taxonomic status/synonym

**Usage**

```
get_status(x, status = "all")
```

**Arguments**

**x** A tibble created with `load_taxonomies()` or `load_population()` or `load_sample()`.

**status** Filter on GBIF assigned status (i.e. NA, "doubtful", "accepted", "proparte synonym", "synonym", "homotypic synonym", "heterotypic synonym"). Can be a string or a vector of strings. Defaults to no filtering.

**Value**

A filtered tibble

**Examples**

```
get_status(load_sample(), "synonym")
get_status(load_sample(), c("accepted", "doubtful"))
```

---

get_validity	<i>Validate entries of a merged taxonomy</i>
--------------	--

---

**Description**

Validate entries of a merged taxonomy

**Usage**

```
get_validity(x, valid = TRUE, rank = "family")
```

**Arguments**

x	A tibble created with <code>load_taxonomies()</code> or <code>load_population()</code> or <code>load_sample()</code> .
valid	A logical indicating whether the returned data should include valid or invalid entries (defaults to <code>valid</code> ).
rank	A string with GBIF rank that will be used to examine an NCBI lineage for validation purposes. Must be "phylum", "class", "order" or "family". Defaults to family.

**Value**

A validated tibble

**Examples**

```
get_validity(load_sample(), valid = TRUE)
```

---

load_population	<i>Load previously merged GBIF and NCBI taxonomies</i>
-----------------	--

---

**Description**

`load_population()` imports a previously merged taxonomy from your file system. An example of an previously merged taxonomy can be downloaded from [https://drive.google.com/file/d/1gpvm9QKd0cuGo\\_cIXPkAgG1B-qfKZZU6/view?usp=sharing](https://drive.google.com/file/d/1gpvm9QKd0cuGo_cIXPkAgG1B-qfKZZU6/view?usp=sharing).

**Usage**

```
load_population(x)
```

**Arguments**

x	Path to a text file containing previously merged NCBI and GBIF taxonomies (compressed or uncompressed).
---	---

**Value**

A tibble

**Examples**

```
## Not run: load_population("path/to/merged_taxonomies")
```

---

load_sample	<i>Load a sample of previously merged GBIF and NCBI taxonomies</i>
-------------	--

---

**Description**

load\_sample() returns a small subset of previously merged GBIF and NCBI taxonomies.

**Usage**

```
load_sample()
```

**Value**

A tibble containing a sample of merged GBIF and NCBI taxonomic data.

**Examples**

```
load_sample()
```

---

load_taxonomies	<i>Load and merge GBIF and NCBI taxonomic data</i>
-----------------	--

---

**Description**

load\_taxonomies() parses and merges a GBIF Taxon.tsv() file (available within the <https://hosted-datasets.gbif.org/datasets/backbone/current/backbone.zip> archive) and a Taxonkit (<https://bioinf.shenwei.me/taxonkit/download/>) output file obtained by running: taxonkit list --ids 1 | taxonkit lineage --show-lineage-taxids --show-lineage-ranks --show-rank --show-name > All.lineages.tsv.

**Usage**

```
load_taxonomies(GBIF_path, NCBI_path)
```

**Arguments**

GBIF_path	Path to the GBIF backbone taxonomy (compressed or uncompressed).
NCBI_path	Path to the NCBI taxonomy (compressed or uncompressed).

**Value**

A tibble containing merged GBIF and NCBI taxonomic data.

**Examples**

```
## Not run: load_taxonomies("path/to/GBIF/Taxon.tsv", "path/to/NCBI-Taxonkit/All.lineages.tsv")
```

plot\_mdb

*Generic for plot\_mdb methods***Description**

A custom generic for methods that plot package:moultdbtools data types

**Usage**

```
plot_mdb(x)
```

**Arguments**

`x` An object of the class `one_rank` or the class `all_ranks`.

**Value**

A ggplot2 derived plot

**Examples**

```
plot_mdb(prepare_rank_dist(load_sample(), NCBI = TRUE, GBIF = TRUE))
plot_mdb(prepare_comparable_rank_dist(load_sample()))
plot_mdb(prepare_rank_dist(get_status(load_sample()), status = "synonym", NCBI = TRUE))
plot_mdb(prepare_comparable_rank_dist(get_validity(get_status(load_sample()), valid = TRUE)))
```

prepare\_comparable\_rank\_dist

*Get comparable NCBI and GBIF taxonomic ranks***Description**

`prepare_comparable_rank_dist()`, like `prepare_rank_dist()`, returns taxonomic ranks aggregated by frequency for data derived from the NCBI, the GBIF, or both. However, `prepare_comparable_rank_dist()` only includes taxonomic ranks that have at least one NCBI and one GBIF representative.

**Usage**

```
prepare_comparable_rank_dist(x, GBIF = TRUE, NCBI = TRUE)
```

**Arguments**

`x` A tibble created with `load_taxonomies()` or `load_population()` or `load_sample()`.  
`GBIF` A boolean indicating whether GBIF taxonomic ranks are to be retrieved.  
`NCBI` A boolean indicating whether NCBI taxonomic ranks are to be retrieved.

**Value**

A list of tibble(s) assigned to the S3 class `one_rank` or to the S3 class `all_ranks`.

**Examples**

```
prepare_comparable_rank_dist(load_sample())
prepare_comparable_rank_dist(get_status(load_sample(), "accepted"), NCBI = FALSE)
```

---

prepare_rank_dist	<i>Get all NCBI and GBIF taxonomic ranks</i>
-------------------	--

---

**Description**

prepare\_rank\_dist() returns taxonomic ranks aggregated by frequency for data derived from the NCBI, the GBIF, or both.

**Usage**

```
prepare_rank_dist(x, GBIF = FALSE, NCBI = FALSE)
```

**Arguments**

x	A tibble created with load_taxonomies() or load_population() or load_sample().
GBIF	A boolean indicating whether GBIF taxonomic ranks are to be retrieved.
NCBI	A boolean indicating whether NCBI taxonomic ranks are to be retrieved.

**Value**

A list of tibble(s) assigned to the S3 class one\_rank or to the S3 class all\_ranks.

**Examples**

```
prepare_rank_dist(load_sample(), NCBI=TRUE, GBIF=TRUE)
prepare_rank_dist(load_sample(), NCBI=TRUE)
```

---

term_conversion	<i>Convert GBIF terms to NCBI terms</i>
-----------------	---

---

**Description**

term\_conversion converts GBIF terminology to NCBI terminology where there is no biological provenance for the difference.

**Usage**

```
term_conversion(x)
```

**Arguments**

x	A tibble created with load_taxonomies() or load_population() or load_sample().
---	--

**Value**

A tibble with converted terms.

**Examples**

```
term_conversion(load_sample())
```

# Index

`get_lineages`, [2](#)  
`get_status`, [2](#)  
`get_validity`, [3](#)  
  
`load_population`, [3](#)  
`load_sample`, [4](#)  
`load_taxonomies`, [4](#)  
  
`plot_mdb`, [5](#)  
`prepare_comparable_rank_dist`, [5](#)  
`prepare_rank_dist`, [6](#)  
  
`term_conversion`, [6](#)