

# Package ‘biomartr’

October 24, 2023

**Title** Genomic Data Retrieval

**Version** 1.0.6

**Description** Perform large scale genomic data retrieval and functional annotation retrieval. This package aims to provide users with a standardized way to automate genome, proteome, 'RNA', coding sequence ('CDS'), 'GFF', and metagenome retrieval from 'NCBI RefSeq', 'NCBI Genbank', 'ENSEMBL', and 'UniProt' databases. Furthermore, an interface to the 'BioMart' database (Smedley et al. (2009) <[doi:10.1186/1471-2164-10-22](https://doi.org/10.1186/1471-2164-10-22)>) allows users to retrieve functional annotation for genomic loci. In addition, users can download entire databases such as 'NCBI RefSeq' (Pruitt et al. (2007) <[doi:10.1093/nar/gkl842](https://doi.org/10.1093/nar/gkl842)>), 'NCBI nr', 'NCBI nt', 'NCBI Genbank' (Ben-son et al. (2013) <[doi:10.1093/nar/gks1195](https://doi.org/10.1093/nar/gks1195)>), etc. with only one command.

**VignetteBuilder** knitr

**NeedsCompilation** yes

**Depends** R (>= 3.1.1)

**Imports** biomaRt, Biostrings, curl, tibble, jsonlite, data.table (>= 1.9.4), dplyr (>= 0.3.0), readr (>= 1.4.0), downloader (>= 0.3), RCurl (>= 1.95-4.5), XML (>= 3.98-1.1), httr (>= 0.6.1), stringr (>= 0.6.2), purrr, R.utils, philentropy, withr, fs

**Suggests** knitr (>= 1.6), rmarkdown (>= 0.3.3), devtools (>= 1.6.1), testthat, seqinr, magrittr

**License** GPL-2

**URL** <https://docs.ropensci.org/biomartr/>,  
<https://github.com/ropensci/biomartr>

**BugReports** <https://github.com/ropensci/biomartr/issues>

**RoxygenNote** 7.2.3

**Encoding** UTF-8

**X-schema.org-keywords** BioMart, genomic-data-retrieval, annotation-retrieval, database-retrieval, NCBI, ENSEMBL, biological-data-retrieval

**X-schema.org-applicationCategory** Data Access

**X-schema.org-isPartof** ``ropensci.org"

**Author** Hajk-Georg Drost [aut, cre] (<<https://orcid.org/0000-0002-1567-306X>>),  
Haakon Tjeldnes [aut, ctb]

**Maintainer** Hajk-Georg Drost <hajk-georg.drost@tuebingen.mpg.de>

**Repository** CRAN

**Date/Publication** 2023-10-24 20:40:06 UTC

## R topics documented:

biomartr-package	3
biomart	4
cachedir	6
cachedir_set	6
check_annotation_biomartr	7
download.database	8
download.database.all	9
ensembl_divisions	10
get.ensembl.info	10
getAssemblyStats	11
getAttributes	13
getBio	14
getBioSet	16
getCDS	18
getCDSSet	20
getCollection	22
getCollectionSet	24
getDatasets	26
getENSEMBL	27
getENSEMBL.gtf	28
getENSEMBL.Seq	29
getENSEMBLGENOMESInfo	30
getENSEMBLInfo	30
getFilters	31
getGenome	32
getGENOMEREPORT	34
getGenomeSet	35
getGFF	37
getGFFSet	39
getGO	41
getGroups	42
getGTF	43
getKingdomAssemblySummary	45
getKingdoms	46
getMarts	47
getMetaGenomeAnnotations	47
getMetaGenomes	48

getMetaGenomeSummary . . . . .	49
getProteome . . . . .	50
getProteomeSet . . . . .	52
getReleases . . . . .	54
getRepeatMasker . . . . .	55
getRNA . . . . .	57
getRNASet . . . . .	59
getSummaryFile . . . . .	61
getUniProtInfo . . . . .	62
getUniProtSTATS . . . . .	62
is.genome.available . . . . .	63
listDatabases . . . . .	64
listGenomes . . . . .	65
listGroups . . . . .	67
listKingdoms . . . . .	68
listMetaGenomes . . . . .	69
meta.retrieval . . . . .	70
meta.retrieval.all . . . . .	73
organismAttributes . . . . .	74
organismBM . . . . .	76
organismFilters . . . . .	77
read_assemblystats . . . . .	79
read_cds . . . . .	79
read_genome . . . . .	81
read_gff . . . . .	82
read_proteome . . . . .	82
read_rm . . . . .	83
read_rna . . . . .	84
refseqOrganisms . . . . .	85
summary_cds . . . . .	85
summary_genome . . . . .	86
<b>Index</b>	<b>88</b>

## Description

This package interacts with a suite of web Application Programming Interfaces and FTP sites to perform automated genomic data retrieval and annotation information retrieval.

## About

To automate the retrieval process on a meta-genomic scale, this package provides useful interface functions for genomic sequence retrieval and functional annotation retrieval. The major aim of `biomartr` is to facilitate computational reproducibility and large-scale handling of genomic data for (meta-)genomic analyses.

In detail, `biomartr` aims to provide users with an easy to use framework to obtain genome, proteome, CDS, GFF (annotation), genome assembly quality, and metagenome project data. Furthermore, an interface to the Ensembl Biomart database allows users to retrieve functional annotation for genomic loci. Users can download entire databases such as

- NCBI RefSeq
- NCBI nr
- NCBI nt
- NCBI Genbank
- NCBI nt
- Ensembl
- Ensembl Genomes
- UniProt

## Author(s)

Hajk-Georg Drost <hajk-georg.drost@tuebingen.mpg.de>

---

biomart

*Main BioMart Query Function*

---

## Description

This function takes a set of gene ids and the biomart specifications and performs a biomart query for the given set of gene ids.

## Usage

```
biomart(genes, mart, dataset, attributes, filters, mute_citation = FALSE, ...)
```

## Arguments

genes	a character vector storing the gene ids of a organisms of interest to be queried against BioMart.
mart	a character string specifying the mart to be used. Users can obtain available marts using <a href="#">getMarts</a> .
dataset	a character string specifying the dataset within the mart to be used, e.g. dataset = "hsapiens_gene_ensembl".

attributes	a character vector specifying the attributes that shall be used, e.g. attributes = c("start_position", "end_position", "description").
filters	a character vector specifying the filter (query key) for the BioMart query, e.g. filter = "ensembl_gene_id".
mute_citation	logical value indicating whether citation message should be muted.
...	additional parameters for the <a href="#">getBM</a> function.

## Details

This function is the main query function of the biomart package.

It enables to fastly access annotations of a given gene set based on the **biomaRt** package implemented by Steffen Durinck et al.

## Value

A data.table storing the initial query gene vector in the first column, the output gene vector in the second column, and all attributes in the following columns.

## Author(s)

Hajk-Georg Drost

## See Also

[organismFilters](#), [organismBM](#), [listAttributes](#), [getBM](#)

## Examples

```
## Not run:
# 1) select a mart
getMarts()

# we will select mart 'plants_mart' and search for available datasets
getDatasets(mart = "plants_mart")

# we choose dataset 'athaliana_eg_gene' and run biomart()
# using mart: 'plants_mart', dataset: "athaliana_eg_gene"
# attributes: c("start_position", "end_position", "description")
# for an example gene set of Arabidopsis thaliana:
# c("AT1G06090", "AT1G06100", "AT1G06110", "AT1G06120",
#   "AT1G06130", "AT1G06200")

biomart(genes      = c("AT1G06090", "AT1G06100",
                      "AT1G06110", "AT1G06120",
                      "AT1G06130", "AT1G06200"),
        mart       = "plants_mart",
        dataset    = "athaliana_eg_gene",
        attributes = c("start_position", "end_position", "description"),
        filters    = "ensembl_gene_id")
```

```
## End(Not run)
```

---

cachedir	<i>Get directory to store back end files like kingdom summaries etc</i>
----------	---

---

**Description**

Get directory to store back end files like kingdom summaries etc

**Usage**

```
cachedir(non_temp_cache = "~/biomartr_cache_dir.rds")
```

**Arguments**

```
non_temp_cache "~/biomartr_cache_dir.rds",
```

**Value**

reads the rds file, and returns the path for local cache, if not existing, use tempdir().

**See Also**

Other cachedir: [cachedir\\_set\(\)](#)

**Examples**

```
cachedir()
```

---

cachedir_set	<i>Set directory to store back end files like kingdom summaries etc</i>
--------------	---

---

**Description**

Set directory to store back end files like kingdom summaries etc

**Usage**

```
cachedir_set(path)
```

**Arguments**

```
path          the path to cache dir, example "~/Bio_data/biomartr_cache/"
```

**Value**

invisible(NULL), only save the file to path location

**See Also**

Other cachedir: [cachedir\(\)](#)

**Examples**

```
# By default it is tempdir()
cachedir()
# cachedir_set("~/Bio_data/biomartr_cache/")
cachedir()
```

---

check\_annotation\_biomartr

*Check whether an annotation file contains outlier lines*

---

**Description**

Some annotation files include lines with character lengths greater than 65000. This causes problems when trying to import such annotation files into R using `import`. To overcome this issue, this function screens for such lines in a given annotation file and removes these lines so that `import` can handle the file.

**Usage**

```
check_annotation_biomartr(annotation_file, remove_annotation_outliers = FALSE)
```

**Arguments**

annotation\_file

a file path to the annotation file.

remove\_annotation\_outliers

shall outlier lines be removed from the input annotation\_file? If yes, then the initial annotation\_file will be overwritten and the removed outlier lines will be stored at [tempdir](#) for further exploration.

**Author(s)**

Hajk-Georg Drost

## Examples

```
## Not run:
# download an example annotation file from NCBI RefSeq
Ath_path <- biomartr::getGFF(organism = "Arabidopsis thaliana")
# run annotation file check on the downloaded file
biomartr::check_annotation_biomartr(Ath_path)
# several outlier lines were detected, thus we re-run the
# function using 'remove_annotation_outliers = TRUE'
# to remove the outliers and overwrite the file
biomartr::check_annotation_biomartr(Ath_path, remove_annotation_outliers = TRUE)

## End(Not run)
```

---

download.database      *Download a NCBI Database to Your Local Hard Drive*

---

## Description

This function allows users to download a database selected by [listDatabases](#) to their local hard drive.

## Usage

```
download.database(db, path = "database")
```

## Arguments

db	a character string specifying the database that shall be downloaded (selected from <a href="#">listDatabases</a> ).
path	a character string specifying the location (a folder) in which the corresponding database shall be stored. Default is path = "database". In case this folder does not exist yet, it will be created.

## Details

This function downloads large databases to your hard drive. For this purpose a folder named database (default) is created and the corresponding database then stored in this folder.

## Value

File path to the downloaded database file.

## Author(s)

Hajk-Georg Drost

## See Also

[download.database.all](#), [listDatabases](#)



## Examples

```
## Not run:
# search for available NCBI nr databases
listNCBIDatabases(db = "nr")
# select NCBI nr version 27 = "nr.27.tar.gz"
# and download it to your hard drive
# -> please note that large databases take some time for download!
download.database(db = "nr.27.tar.gz")

## End(Not run)
```

---

download.database.all *Download all elements of an NCBI databse*

---

## Description

The [download.database](#) functions allows users to retrieve individual packages of a NCBI database. This function is designed to retrieve the entire database selected by the users (hence all packages corresponding to this database).

## Usage

```
download.database.all(db, path = NULL)
```

## Arguments

db	a character string specifying the database that shall be downloaded (selected from <a href="#">listDatabases</a> ).
path	a character string specifying the location (a folder) in which the corresponding database shall be stored. In case this folder does not exist yet, it will be created.

## Value

A character vector storing the file paths of the downloaded databases.

## Author(s)

Hajk-Georg Drost

## See Also

[download.database](#), [listNCBIDatabases](#)

**Examples**

```
## Not run:
# search for available NCBI databases
listNCBIDatabases(db = "all")
# choose database NCBI nr and download complete database
download.database.all(db = "nr", path = "nr")

## End(Not run)
```

---

ensembl_divisions	<i>List all available ENSEMBL divisions</i>
-------------------	---

---

**Description**

Retrieve a list of available databases on ENSEMBL for which `get.ensembl.info` can be retrieved.

**Usage**

```
ensembl_divisions()
```

**Author(s)**

Hajk-Georg Drost

**Examples**

```
ensembl_divisions()
```

---

get.ensembl.info	<i>Helper function to retrieve species information from the ENSEMBL API</i>
------------------	---

---

**Description**

This function interfaces with the ENSEMBL API (<https://rest.ensembl.org/info/species?content-type=application/json>) and internally stores the output to use this information for subsequent retrieval function calls.

**Usage**

```
get.ensembl.info(update = FALSE, division)
```

**Arguments**

update	logical, default FALSE. If TRUE, force re-download of info.
division	the ENSEMBL database (division) for which information shall be retrieved (available options can be obtained with <code>ensembl_divisions</code> ).

**Author(s)**

Hajk-Georg Drost

**See Also**[ensembl\\_divisions](#), [getKingdomAssemblySummary](#), [getENSEMBLInfo](#)**Examples**

```
## Not run:
# Look at available ENSEMBL division options
ensembl_divisions()
# Retrieve available information for EnsemblVertebrates
example <- get.ensembl.info(division = "EnsemblVertebrates")
example
# Update information file stored in the tempdir() folder.
example_update <- get.ensembl.info(division = "EnsemblVertebrates", update = TRUE)
example_update

## End(Not run)
```

---

`getAssemblyStats`*Genome Assembly Stats Retrieval*

---

**Description**

Main genome assembly stats retrieval function for an organism of interest. By specifying the scientific name of an organism of interest the corresponding genome assembly stats file storing the assembly statistics of the organism of interest can be downloaded and stored locally. Genome assembly stats files can be retrieved from several databases.

**Usage**

```
getAssemblyStats(
  db = "refseq",
  organism,
  reference = FALSE,
  skip_bacteria = TRUE,
  release = NULL,
  type = "download",
  path = file.path("_ncbi_downloads", "genomeassembly_stats"),
  mute_citation = FALSE
)
```

**Arguments**

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organism	a character string specifying the scientific name of the organism of interest, e.g. organism = "Homo sapiens".
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
release	most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data in the standard format before that.
type	shall only the file be retrieved (default) type = "download" or should the corresponding file be downloaded and subsequently be imported type = "import".
path	a character string specifying the location (a folder) in which the corresponding file shall be stored. Default is path = file.path("_ncbi_downloads", "genomeassembly_stats").
mute_citation	logical value indicating whether citation message should be muted.

**Details**

Internally this function loads the the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

to retrieve available scientific names of organisms and creates a directory '\_ncbi\_downloads/genomeassembly\_stats' to store the Genome Assembly Stats of interest as text file for future processing. In case the corresponding fasta file already exists within the '\_ncbi\_downloads/genomeassembly\_stats' folder and is accessible within the workspace, no download process will be performed.

An example genome assembly stats file can be found here: [ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF\\_000001405.36\\_GRCh38.p10/GCF\\_000001405.36\\_GRCh38.p10\\_assembly\\_stats.txt](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.36_GRCh38.p10/GCF_000001405.36_GRCh38.p10_assembly_stats.txt).

**Value**

File path to downloaded genome assembly stats file.

**Author(s)**

Hajk-Georg Drost

**See Also**

[getGenome](#), [getProteome](#), [getCDS](#), [getGFF](#), [getRNA](#), [getCollection](#), [meta.retrieval](#), [read\\_assemblystats](#)

**Examples**

```
## Not run:
# download the genome assembly stats file of Saccharomyces cerevisiae
# from NCBI RefSeq
# and store the corresponding genome file in
# '_ncbi_downloads/genomeassembly_stats'
file_path <- getAssemblyStats( db = "refseq",
                              organism = "Saccharomyces cerevisiae",
                              path = file.path("_ncbi_downloads", "genomeassembly_stats"))
# import the raw file as it is downloaded
Scerevisiae.stats <- read_assemblystats(file_path, type = "raw")

# download the genome assembly stats file of Saccharomyces cerevisiae
# from NCBI RefSeq
# and import overall statistics of the genome assembly
Scerevisiae.stats.import <- getAssemblyStats( db = "refseq",
                                              organism = "Saccharomyces cerevisiae",
                                              type = "import",
                                              path = file.path("_ncbi_downloads", "genomeassembly_stats"))

## End(Not run)
```

---

getAttributes

*Retrieve All Available Attributes for a Specific Dataset*

---

**Description**

This function queries the BioMart Interface and returns a table storing all available attributes for a specific dataset.

**Usage**

```
getAttributes(mart, dataset, mute_citation = FALSE)
```

**Arguments**

mart	a character string specifying the database (mart) for which datasets shall be listed.
dataset	a character string specifying the dataset for which attributes shall be listed.
mute_citation	logical value indicating whether citation message should be muted.

**Author(s)**

Hajk-Georg Drost

**See Also**

[getMarts](#), [getDatasets](#), [getFilters](#), [organismBM](#), [organismFilters](#), [organismAttributes](#)

**Examples**

```
## Not run:
# search for available datasets
getMarts()

# choose database (mart): ENSEMBL_MART_ENSEMBL
# and get a table of all available datasets from this BioMart database
head(getDatasets(mart = "ENSEMBL_MART_ENSEMBL"), 10)

# choose dataset: "hsapiens_gene_ensembl"
head(getAttributes(mart = "ENSEMBL_MART_ENSEMBL",
                  dataset = "hsapiens_gene_ensembl") , 5)

## End(Not run)
```

---

getBio

*A wrapper to all bio getters, selected with 'type' argument*

---

**Description**

A wrapper to all bio getters, selected with 'type' argument

**Usage**

```
getBio(
  db = "refseq",
  organism,
  type,
  reference = FALSE,
  release = NULL,
  gunzip = FALSE,
  update = FALSE,
  skip_bacteria = TRUE,
  path = paste0("set_", toupper(type)),
  remove_annotation_outliers = FALSE,
  analyse_genome = FALSE,
  assembly_type = "toplevel",
  format = "gff3",
  mute_citation = FALSE
)
```

**Arguments**

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organism	Organism selector id, there are three options to characterize an organism: <ul style="list-style-type: none"> <li>• by scientific name: e.g. organism = "Homo sapiens"</li> <li>• by database specific accession identifier: e.g. organism = "GCF_000001405.37" (= NCBI RefSeq identifier for Homo sapiens)</li> <li>• by taxonomic identifier from NCBI Taxonomy: e.g. organism = "9606" (= taxid of Homo sapiens)</li> </ul>
type	biological sequence type. (alternatives are: genome, gff, cds, rna, proteome, assembly_stats, repeat_masker, collection (all the others))
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data in the standard format before that.
gunzip	a logical, indicating whether or not files should be unzipped.
update	logical, default FALSE. Updated backend cached files needed. Usually keep this false, to make it run much faster. Only set to TRUE, if you believe your cache is outdated (Species only exist in newest release etc)
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
path	character, default location is paste0("set_", toupper(type))
remove_annotation_outliers	shall outlier lines be removed from the input annotation_file? If yes, then the initial annotation_file will be overwritten and the removed outlier lines will be stored at <code>tempdir</code> for further exploration.
analyse_genome	logical, default FALSE. If TRUE, get general genome statistics like gc content etc. For more details, see <code>?summary_genome</code>
assembly_type	a character, default "toplevel". id type of assembly, either "toplevel" or "primary_assembly" usually.
format	"gff3", alternative "gtf" for ensembl.
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

## Details

Internally this function loads the the overview.txt file from NCBI:

refseq: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/

genbank: ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/

and creates a directory relative to file type, if you get fasta genomes it will be `_ncbi_downloads/genomes'`.

In case the corresponding fasta file already exists within the `'_ncbi_downloads/genomes'` folder and is accessible within the workspace, no download process will be performed. For other file types the same rule applies.

## Value

File path to downloaded genome.

## Author(s)

Hajk-Georg Drost

## See Also

Other getBio: [getCDS\(\)](#), [getCollection\(\)](#), [getGFF\(\)](#), [getGenome\(\)](#), [getProteome\(\)](#), [getRNA\(\)](#)

---

getBioSet

*Generic Bio data set extractor*

---

## Description

Usually you want to use one of the specific set extractors

## Usage

```
getBioSet(  
  db = "refseq",  
  organisms,  
  set_type,  
  reference = FALSE,  
  release = NULL,  
  gunzip = TRUE,  
  update = FALSE,  
  skip_bacteria = TRUE,  
  path = paste0("set_", toupper(set_type)),  
  remove_annotation_outliers = FALSE,  
  assembly_type = "toplevel",  
  format = "gff3",  
  mute_citation = FALSE  
)
```



**Arguments**

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organisms	a character vector storing the names of the organisms than shall be retrieved. There are three available options to characterize an organism:
set_type	the biological sequence type that shall be retrieved. Available options are <ul style="list-style-type: none"> <li>• set_type = "genome"</li> <li>• set_type = "proteome"</li> <li>• set_type = "cds"</li> <li>• set_type = "gff"</li> <li>• set_type = "rna"</li> <li>• set_type = "assembly_stats"</li> <li>• set_type = "repeat_masker"</li> <li>• set_type = "collection" (all the others)</li> </ul>
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
gunzip	a logical, indicating whether or not files should be unzipped.
update	logical, default FALSE. Updated backend cached files needed. Usually keep this false, to make ut run much faster. Only set to TRUE, if you believe you cache is outdated (Species only exist in newest release etc)
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they needs to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
path	character, default location is paste0("set_", toupper(set_type))
remove_annotation_outliers	shall outlier lines be removed from the input annotation_file? If yes, then the initial annotation_file will be overwritten and the removed outlier lines will be stored at <a href="#">tempdir</a> for further exploration.
assembly_type	a character, default "toplevel". id type of assembly, either "toplevel" or "primary_assembly" usually.
format	"gff3", alternative "gtf" for ensembl.
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

## Details

Internally this function loads the the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

and creates a directory 'set\_CDSs' to store the CDSs of interest as fasta files for future processing. In case the corresponding fasta file already exists within the 'set\_CDSs' folder and is accessible within the workspace, no download process will be performed.

## Value

File path to downloaded genomes (names are identifiers: 'new' (file was downloaded now), 'old' files did already exist)

## Author(s)

Hajk-Georg Drost

## See Also

[getBio](#)

Other `getBioSet`: [getCDSSet\(\)](#), [getCollectionSet\(\)](#), [getGFFSet\(\)](#), [getGenomeSet\(\)](#), [getProteomeSet\(\)](#), [getRNASet\(\)](#)

## Examples

```
## Not run:
getBioSet("refseq", organisms = c("Arabidopsis thaliana",
                                  "Arabidopsis lyrata",
                                  "Capsella rubella"),
          set_type = "cds")

## End(Not run)
```

---

getCDS

*Coding Sequence Retrieval*

---

## Description

Main retrieval function for coding sequences (CDS) of an organism of interest. By specifying the scientific name of an organism of interest the corresponding fasta-file storing the CDS information for the organism of interest can be downloaded and stored locally. CDS files can be retrieved from several databases.

**Usage**

```

getCDS(
  db = "refseq",
  organism,
  reference = FALSE,
  skip_bacteria = TRUE,
  release = NULL,
  gunzip = FALSE,
  path = file.path("_ncbi_downloads", "CDS"),
  mute_citation = FALSE
)

```

**Arguments**

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organism	Organism selector id, there are three options to characterize an organism: <ul style="list-style-type: none"> <li>• by scientific name: e.g. organism = "Homo sapiens"</li> <li>• by database specific accession identifier: e.g. organism = "GCF_000001405.37" (= NCBI RefSeq identifier for Homo sapiens)</li> <li>• by taxonomic identifier from NCBI Taxonomy: e.g. organism = "9606" (= taxid of Homo sapiens)</li> </ul>
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data in the standard format before that.
gunzip	a logical, indicating whether or not files should be unzipped.
path	a character string specifying the location (a folder) in which the corresponding CDS file shall be stored. Default is path = file.path("_ncbi_downloads", "CDS").
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

**Details**

Internally this function loads the overview.txt file from NCBI:

refseq: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/

genbank: ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/

and creates a directory relative to file type, if you get fasta genomes it will be `_ncbi_downloads/genomes/`. In case the corresponding fasta file already exists within the `'_ncbi_downloads/genomes'` folder and is accessible within the workspace, no download process will be performed. For other file types the same rule applies.

### Value

File path to downloaded genome.

### Author(s)

Hajk-Georg Drost

### See Also

Other getBio: [getBio\(\)](#), [getCollection\(\)](#), [getGFF\(\)](#), [getGenome\(\)](#), [getProteome\(\)](#), [getRNA\(\)](#)

Other cds: [getCDSset\(\)](#), [read\\_cds\(\)](#)

### Examples

```
## Not run:
# download the genome of Arabidopsis thaliana from refseq
# and store the corresponding genome CDS file in '_ncbi_downloads/CDS'
file_path <- getCDS( db      = "refseq",
                    organism = "Arabidopsis thaliana",
                    path    = file.path("_ncbi_downloads", "CDS"))

Ath_CDS <- read_cds(file_path, format = "fasta")

## End(Not run)
```

---

getCDSset

*CDS retrieval of multiple species*

---

### Description

Main CDS retrieval function for a set of organism of interest. By specifying the scientific names of the organisms of interest the corresponding fasta-files storing the CDS of the organisms of interest will be downloaded and stored locally. CDS files can be retrieved from several databases.

**Usage**

```

getCDSset(
  db = "refseq",
  organisms,
  reference = FALSE,
  release = NULL,
  gunzip = TRUE,
  update = FALSE,
  path = "set_CDS"
)

```

**Arguments**

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organisms	a character vector storing the names of the organisms than shall be retrieved. There are three available options to characterize an organism:
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
gunzip	a logical, indicating whether or not files should be unzipped.
update	logical, default FALSE. Updated backend cached files needed. Usually keep this false, to make ut run much faster. Only set to TRUE, if you believe you cache is outdated (Species only exist in newest release etc)
path	character, default location is paste0("set_", toupper(set_type))

**Details**

Internally this function loads the the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

and creates a directory 'set\_CDSs' to store the CDSs of interest as fasta files for future processing. In case the corresponding fasta file already exists within the 'set\_CDSs' folder and is accessible within the workspace, no download process will be performed.

**Value**

File path to downloaded genomes (names are identifiers: 'new' (file was downloaded now), 'old' files did already exist)

**Author(s)**

Hajk-Georg Drost

**See Also**

Other getBioSet: [getBioSet\(\)](#), [getCollectionSet\(\)](#), [getGFFSet\(\)](#), [getGenomeSet\(\)](#), [getProteomeSet\(\)](#), [getRNASet\(\)](#)

Other cds: [getCDS\(\)](#), [read\\_cds\(\)](#)

**Examples**

```
## Not run:
getBioSet("refseq", organisms = c("Arabidopsis thaliana",
                                "Arabidopsis lyrata",
                                "Capsella rubella"),
         set_type = "cds")

## End(Not run)
```

---

getCollection	<i>Retrieve a Collection: Genome, Proteome, CDS, RNA, GFF, Repeat Masker, AssemblyStats</i>
---------------	---

---

**Description**

Main collection retrieval function for an organism of interest. By specifying the scientific name of an organism of interest a collection consisting of the genome file, proteome file, CDS file, RNA file, GFF file, Repeat Masker file, AssemblyStats file of the organism of interest can be downloaded and stored locally. Collections can be retrieved from several databases.

**Usage**

```
getCollection(
  db = "refseq",
  organism,
  reference = TRUE,
  skip_bacteria = TRUE,
  release = NULL,
  assembly_type = "toplevel",
  analyse_genome = FALSE,
  remove_annotation_outliers = FALSE,
  gunzip = FALSE,
  path = file.path("_db_downloads", "collections"),
  mute_citation = FALSE
)
```

**Arguments**

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organism	Organism selector id, there are three options to characterize an organism: <ul style="list-style-type: none"> <li>• by scientific name: e.g. organism = "Homo sapiens"</li> <li>• by database specific accession identifier: e.g. organism = "GCF_000001405.37" (= NCBI RefSeq identifier for Homo sapiens)</li> <li>• by taxonomic identifier from NCBI Taxonomy: e.g. organism = "9606" (= taxid of Homo sapiens)</li> </ul>
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data in the standard format before that.
assembly_type	a character, default "toplevel". id type of assembly, either "toplevel" or "primary_assembly" usually.
analyse_genome	logical, default FALSE. If TRUE, get general genome statistics like gc content etc. For more details, see ?summary_genome
remove_annotation_outliers	shall outlier lines be removed from the input annotation_file? If yes, then the initial annotation_file will be overwritten and the removed outlier lines will be stored at <a href="#">tempdir</a> for further exploration.
gunzip	a logical, indicating whether or not files should be unzipped.
path	a character string specifying the location (a folder) in which the corresponding collection shall be stored. Default is path = file.path("_db_downloads", "collections").
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

**Details**

Internally this function loads the the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

and creates a directory relative to file type, if you get fasta genomes it will be `_ncbi_downloads/genomes`. In case the corresponding fasta file already exists within the `'_ncbi_downloads/genomes'` folder and is accessible within the workspace, no download process will be performed. For other file types the same rule applies.

### Value

File path to downloaded genome.

### Author(s)

Hajk-Georg Drost

### See Also

Other getBio: [getBio\(\)](#), [getCDS\(\)](#), [getGFF\(\)](#), [getGenome\(\)](#), [getProteome\(\)](#), [getRNA\(\)](#)

Other collection: [getCollectionSet\(\)](#)

### Examples

```
## Not run:
# download the collection of Homo sapiens from refseq
# and store the corresponding genome file in '_ncbi_downloads/collection'
Hsap_collection <- getCollection( db      = "refseq",
                                organism = "Homo sapiens",
                                path = file.path("_db_downloads", "collections"))
# download the collection of Homo sapiens from genbank
# and store the corresponding genome file in '_ncbi_downloads/collection'
Hsap_collection <- getCollection( db      = "genbank",
                                organism = "Homo sapiens",
                                path = file.path("_db_downloads", "collections"))
# download the collection of Homo sapiens from ensembl
# and store the corresponding genome file in '_ncbi_downloads/collection'
Hsap_collection <- getCollection( db      = "ensembl",
                                organism = "Homo sapiens",
                                path = file.path("_db_downloads", "collections"))

## End(Not run)
```

---

<code>getCollectionSet</code>	<i>Retrieve a Collection: Genome, Proteome, CDS, RNA, GFF, Repeat Masker, AssemblyStats of multiple species</i>
-------------------------------	---

---

### Description

Main collection retrieval function for an organism of interest. By specifying the scientific name of an organism of interest a collection consisting of the genome file, proteome file, CDS file, RNA file, GFF file, Repeat Masker file, AssemblyStats file of the organism of interest can be downloaded and stored locally. Collections can be retrieved from several databases.



**Usage**

```

getCollectionSet(
  db = "refseq",
  organisms,
  reference = FALSE,
  release = NULL,
  skip_bacteria = TRUE,
  gunzip = TRUE,
  update = FALSE,
  remove_annotation_outliers = TRUE,
  path = "set_collections",
  mute_citation = FALSE
)

```

**Arguments**

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organisms	a character vector storing the names of the organisms than shall be retrieved. There are three available options to characterize an organism:
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they needs to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
gunzip	a logical, indicating whether or not files should be unzipped.
update	logical, default FALSE. Updated backend cached files needed. Usually keep this false, to make ut run much faster. Only set to TRUE, if you believe you cache is outdated (Species only exist in newest release etc)
remove_annotation_outliers	shall outlier lines be removed from the input annotation_file? If yes, then the initial annotation_file will be overwritten and the removed outlier lines will be stored at <a href="#">tempdir</a> for further exploration.
path	a character string specifying the location (a folder) in which the corresponding collection shall be stored. Default is path = file.path("_db_downloads", "collections").
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

## Details

Internally this function loads the the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

and creates a directory 'set\_CDSs' to store the CDSs of interest as fasta files for future processing. In case the corresponding fasta file already exists within the 'set\_CDSs' folder and is accessible within the workspace, no download process will be performed.

## Value

File path to downloaded genomes (names are identifiers: 'new' (file was downloaded now), 'old' files did already exist)

## Author(s)

Hajk-Georg Drost

## See Also

Other getBioSet: [getBioSet\(\)](#), [getCDSSet\(\)](#), [getGFFSet\(\)](#), [getGenomeSet\(\)](#), [getProteomeSet\(\)](#), [getRNASet\(\)](#)

Other collection: [getCollection\(\)](#)

## Examples

```
## Not run:
getBioSet("refseq", organisms = c("Arabidopsis thaliana",
                                  "Arabidopsis lyrata",
                                  "Capsella rubella"),
          set_type = "cds")

## End(Not run)
```

---

getDatasets

*Retrieve All Available Datasets for a BioMart Database*

---

## Description

This function queries the BioMart API and returns a table storing all available datasets for a selected BioMart databases.

## Usage

```
getDatasets(mart, mute_citation = FALSE)
```

**Arguments**

- `mart` a character string specifying the database (mart) for which datasets shall be listed.
- `mute_citation` logical value indicating whether citation message should be muted.

**Author(s)**

Hajk-Georg Drost

**See Also**

[getMarts](#), [getAttributes](#), [getFilters](#), [organismBM](#), [organismFilters](#), [organismAttributes](#)

**Examples**

```
## Not run:  
# search for available datasets  
# getMarts()  
# choose database: "ENSEMBL_MART_ENSEMBL"  
head(getDatasets("ENSEMBL_MART_ENSEMBL"), 10)  
  
## End(Not run)
```

---

getENSEMBL

*Download sequence or annotation from ENSEMBL*

---

**Description**

Backend function for retrieving files sequence and annotation files from the ENSEMBL ftp server

**Usage**

```
getENSEMBL(  
  organism,  
  type = "dna",  
  id.type = "toplevel",  
  release = NULL,  
  path,  
  format  
)
```

**Arguments**

- `organism` Organism selector id, there are three options to characterize an organism:
- by scientific name: e.g. `organism = "Homo sapiens"`
  - by database specific accession identifier: e.g. `organism = "GCF_000001405.37"` (= NCBI RefSeq identifier for Homo sapiens)

	<ul style="list-style-type: none"> <li>• by taxonomic identifier from NCBI Taxonomy: e.g. organism = "9606" (= taxid of Homo sapiens)</li> </ul>
type	character, biological sequence type (e.g. "dna", "cds")
id.type	a character, default "toplevel". id type of assembly, either "toplevel" or "primary_assembly" usually.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
path	location where file shall be stored.
format	"gff3", alternative "gtf" for ensembl.

**Value**

either a character path to downloaded file, or a logical FALSE, specifying failure.

**Author(s)**

Hajk-Georg Drost

---

getENSEMBL.gtf

*Helper function for retrieving gtf files from ENSEMBL*

---

**Description**

This function downloads gff files of query organisms from ENSEMBL.

**Usage**

```
getENSEMBL.gtf(organism, type = "dna", path, release = NULL)
```

**Arguments**

organism	<p>Organism selector id, there are three options to characterize an organism:</p> <ul style="list-style-type: none"> <li>• by scientific name: e.g. organism = "Homo sapiens"</li> <li>• by database specific accession identifier: e.g. organism = "GCF_000001405.37" (= NCBI RefSeq identifier for Homo sapiens)</li> <li>• by taxonomic identifier from NCBI Taxonomy: e.g. organism = "9606" (= taxid of Homo sapiens)</li> </ul>
type	character, biological sequence type (e.g. "dna", "cds")
path	location where file shall be stored.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.

**Value**

character filepath to download file, returns FALSE if failed.

**Author(s)**

Hajk-Georg Drost

---

getENSEMBL.Seq	<i>Helper function for retrieving biological sequence files from ENSEMBL</i>
----------------	--

---

**Description**

This function downloads gff files of query organisms from ENSEMBL.

**Usage**

```
getENSEMBL.Seq(
  organism,
  type = "dna",
  id.type = "toplevel",
  release = NULL,
  path
)
```

**Arguments**

organism	Organism selector id, there are three options to characterize an organism: <ul style="list-style-type: none"> <li>• by scientific name: e.g. organism = "Homo sapiens"</li> <li>• by database specific accession identifier: e.g. organism = "GCF_000001405.37" (= NCBI RefSeq identifier for Homo sapiens)</li> <li>• by taxonomic identifier from NCBI Taxonomy: e.g. organism = "9606" (= taxid of Homo sapiens)</li> </ul>
type	character, biological sequence type (e.g. "dna", "cds")
id.type	a character, default "toplevel". id type of assembly, either "toplevel" or "primary_assembly" usually.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
path	location where file shall be stored.

**Value**

either a character path to downloaded file, or a logical FALSE, specifying failure.

**Author(s)**

Hajk-Georg Drost

---

getENSEMBLGENOMESInfo *Retrieve ENSEMBLGENOMES info file*

---

**Description**

Retrieve species and genome information from <http://rest.ensemblgenomes.org/info/species?content-type=application/json/>.

**Usage**

```
getENSEMBLGENOMESInfo()
```

**Author(s)**

Hajk-Georg Drost

**Examples**

```
## Not run:  
info.file <- getENSEMBLGENOMESInfo()  
info.file  
  
## End(Not run)
```

---

getENSEMBLInfo *Retrieve ENSEMBL info file*

---

**Description**

Retrieve species and genome information from <http://rest.ensembl.org/info/species?content-type=application/json/>.

**Usage**

```
getENSEMBLInfo(update = TRUE)
```

**Arguments**

update           logical, default TRUE. Update cached list, if FALSE use existing (if it exists)

**Value**

a tibble table storing info for all available ENSEMBL divisions.

**Author(s)**

Hajk-Georg Drost

**See Also**[ensembl\\_divisions](#), [get.ensembl.info](#), [getKingdomAssemblySummary](#)**Examples**

```
## Not run:
# look at available divisions
ensembl_divisions()
# retrieve information for all ENSEMBL divisions at once
test <- getENSEMBLInfo()
test
# retrieve information for a particular ENSEMBL division (e.g. EnsemblVertebrates)
test_vertebrates <- get.ensembl.info(update = TRUE, division = "EnsemblVertebrates")
test_vertebrates

## End(Not run)
```

---

`getFilters`*Retrieve All Available Filters for a Specific Dataset*

---

**Description**

This function queries the BioMart API and returns a table storing all available filters for a specific dataset.

**Usage**

```
getFilters(mart, dataset, mute_citation = FALSE)
```

**Arguments**

<code>mart</code>	a character string specifying the database (mart) for which datasets shall be listed.
<code>dataset</code>	a character string specifying the dataset for which filters shall be listed.
<code>mute_citation</code>	logical value indicating whether citation message should be muted.

**Author(s)**

Hajk-Georg Drost

**See Also**[getMarts](#), [getDatasets](#), [getAttributes](#), [organismBM](#), [organismFilters](#), [organismAttributes](#)

## Examples

```
## Not run:
# search for available datasets
# getMarts()
# choose database (mart): "ENSEMBL_MART_ENSEMBL"
# head(getDatasets(mart = "ENSEMBL_MART_ENSEMBL"), 10)
# choose dataset: "hsapiens_gene_ensembl"
head(getFilters(mart = "ENSEMBL_MART_ENSEMBL",
                dataset = "hsapiens_gene_ensembl"), 5)

## End(Not run)
```

---

getGenome

*Genome Retrieval*

---

## Description

Main genome retrieval function for an organism of interest. By specifying the scientific name of an organism of interest the corresponding fasta-file storing the genome of the organism of interest can be downloaded and stored locally. Genome files can be retrieved from several databases. In addition, the genome summary statistics for the retrieved species is stored locally to provide users with insights regarding the genome assembly quality (see [summary\\_genome](#) for details). This is useful when comparing genomes with large difference in genome assembly qualities.

## Usage

```
getGenome(
  db = "refseq",
  organism,
  reference = FALSE,
  skip_bacteria = TRUE,
  release = NULL,
  gunzip = FALSE,
  path = file.path("_ncbi_downloads", "genomes"),
  assembly_type = "toplevel",
  mute_citation = FALSE,
  analyse_genome = FALSE
)
```

## Arguments

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organism	Organism selector id, there are three options to characterize an organism:



- by scientific name: e.g. organism = "Homo sapiens"
- by database specific accession identifier: e.g. organism = "GCF\_000001405.37" (= NCBI RefSeq identifier for Homo sapiens)
- by taxonomic identifier from NCBI Taxonomy: e.g. organism = "9606" (= taxid of Homo sapiens)

reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data in the standard format before that.
gunzip	a logical, indicating whether or not files should be unzipped.
path	character, default location is paste0("set_", toupper(type))
assembly_type	a character, default "toplevel". id type of assembly, either "toplevel" or "primary_assembly" usually.
mute_citation	logical, default FALSE, indicating whether citation message should be muted.
analyse_genome	logical, default FALSE. If TRUE, get general genome statistics like gc content etc. For more details, see ?summary_genome

### Details

Internally this function loads the the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

and creates a directory relative to file type, if you get fasta genomes it will be `_ncbi_downloads/genomes'`. In case the corresponding fasta file already exists within the `'_ncbi_downloads/genomes'` folder and is accessible within the workspace, no download process will be performed. For other file types the same rule applies.

### Value

File path to downloaded genome.

### Author(s)

Hajk-Georg Drost

**See Also**

Other getBio: [getBio\(\)](#), [getCDS\(\)](#), [getCollection\(\)](#), [getGFF\(\)](#), [getProteome\(\)](#), [getRNA\(\)](#)

Other genome: [getGenomeSet\(\)](#), [read\\_genome\(\)](#)

---

`getGENOMEREPORT`*Retrieve NCBI GENOME\_REPORTS file*

---

**Description**

Retrieves NCBI GENOME\_REPORTS file from [ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME\\_REPORTS/overview.txt](ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/overview.txt).

**Usage**

```
getGENOMEREPORT(  
  local_file = file.path(cachedir(), "_ncbi_downloads", "overview.txt")  
)
```

**Arguments**

`local_file` character, file path, default: `file.path(cachedir(), "_ncbi_downloads", "overview.txt")`

**Value**

a tibble object with the report

**Author(s)**

Hajk-Georg Drost

**Examples**

```
## Not run:  
report <- getGENOMEREPORT()  
report  
  
## End(Not run)
```

---

getGenomeSet

*Genome Retrieval of multiple species*


---

### Description

Main genome retrieval function for a set of organism of interest. By specifying the scientific names of the organisms of interest the corresponding fasta-files storing the genome of the organisms of interest will be downloaded and stored locally. Genome files can be retrieved from several databases.

### Usage

```
getGenomeSet(
  db = "refseq",
  organisms,
  reference = FALSE,
  release = NULL,
  skip_bacteria = TRUE,
  gunzip = TRUE,
  update = FALSE,
  path = "set_genomes",
  assembly_type = "toplevel",
  mute_citation = FALSE
)
```

### Arguments

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organisms	a character vector storing the names of the organisms than shall be retrieved. There are three available options to characterize an organism:
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they needs to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.

gunzip	a logical, indicating whether or not files should be unzipped.
update	logical, default FALSE. Updated backend cached files needed. Usually keep this false, to make it run much faster. Only set to TRUE, if you believe your cache is outdated (Species only exist in newest release etc)
path	a character string specifying the location (a folder) in which the corresponding genomes shall be stored. Default is path = "set_genomes".
assembly_type	a character, default "toplevel". id type of assembly, either "toplevel" or "primary_assembly" usually.
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

### Details

Internally this function loads the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

and creates a directory 'set\_CDSs' to store the CDSs of interest as fasta files for future processing. In case the corresponding fasta file already exists within the 'set\_CDSs' folder and is accessible within the workspace, no download process will be performed.

### Value

File path to downloaded genomes (names are identifiers: 'new' (file was downloaded now), 'old' files did already exist)

### Author(s)

Hajk-Georg Drost

### See Also

Other getBioSet: [getBioSet\(\)](#), [getCDSSet\(\)](#), [getCollectionSet\(\)](#), [getGFFSet\(\)](#), [getProteomeSet\(\)](#), [getRNASet\(\)](#)

Other genome: [getGenome\(\)](#), [read\\_genome\(\)](#)

### Examples

```
## Not run:
getBioSet("refseq", organisms = c("Arabidopsis thaliana",
                                "Arabidopsis lyrata",
                                "Capsella rubella"),
         set_type = "cds")
```

```
## End(Not run)
```

getGFF

*Genome Annotation Retrieval (GFF3)*

## Description

Main retrieval function for GFF files of an organism of interest. By specifying the scientific name of an organism of interest the corresponding gff file storing the annotation for the organism of interest can be downloaded and stored locally. GFF files can be retrieved from several databases.

## Usage

```
getGFF(
  db = "refseq",
  organism,
  reference = FALSE,
  skip_bacteria = TRUE,
  release = NULL,
  gunzip = FALSE,
  remove_annotation_outliers = FALSE,
  path = file.path("_ncbi_downloads", "annotation"),
  mute_citation = FALSE,
  format = "gff3"
)
```

## Arguments

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organism	Organism selector id, there are three options to characterize an organism: <ul style="list-style-type: none"> <li>• by scientific name: e.g. organism = "Homo sapiens"</li> <li>• by database specific accession identifier: e.g. organism = "GCF_000001405.37" (= NCBI RefSeq identifier for Homo sapiens)</li> <li>• by taxonomic identifier from NCBI Taxonomy: e.g. organism = "9606" (= taxid of Homo sapiens)</li> </ul>
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.

release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
gunzip	a logical, indicating whether or not files should be unzipped.
remove_annotation_outliers	shall outlier lines be removed from the input annotation_file? If yes, then the initial annotation_file will be overwritten and the removed outlier lines will be stored at <code>tempdir</code> for further exploration.
path	a character string specifying the location (a folder) in which the corresponding annotation file shall be stored. Default is path = <code>file.path("_ncbi_downloads", "annotation")</code> .
mute_citation	logical, default FALSE, indicating whether citation message should be muted.
format	"gff3", alternative "gtf" for ensembl.

### Details

Internally this function loads the the overview.txt file from NCBI:

refseq: `ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/`

genbank: `ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/`

and creates a directory relative to file type, if you get fasta genomes it will be `_ncbi_downloads/genomes'`. In case the corresponding fasta file already exists within the `'_ncbi_downloads/genomes'` folder and is accessible within the workspace, no download process will be performed. For other file types the same rule applies.

### Value

File path to downloaded genome.

### Author(s)

Hajk-Georg Drost

### See Also

Other getBio: `getBio()`, `getCDS()`, `getCollection()`, `getGenome()`, `getProteome()`, `getRNA()`

Other gff: `getGFFSet()`, `read_gff()`

### Examples

```
## Not run:
# download the annotation of Arabidopsis thaliana from refseq
# and store the corresponding genome file in '_ncbi_downloads/annotation'
Athal_gff <- getGFF( db      = "refseq",
                   organism = "Arabidopsis thaliana",
                   path = file.path("_ncbi_downloads", "annotation"),
                   remove_annotation_outliers = TRUE)
```

```

Athal_gff_import <- read_gff(Athal_gff)

# download the genome of Arabidopsis thaliana from genbank
# and store the corresponding genome file in '_ncbi_downloads/annotation'
Athal_gff <- getGFF( db      = "genbank",
                    organism = "Arabidopsis thaliana",
                    path = file.path("_ncbi_downloads","annotation"),
                    remove_annotation_outliers = TRUE)
Athal_gff_import <- read_gff(Athal_gff)

# download the genome of Homo sapiens from ensembl
# and store the corresponding genome file in '_ncbi_downloads/annotation'
Hsap_gff <- getGFF( db      = "ensembl",
                   organism = "Homo sapiens",
                   path = file.path("_ncbi_downloads","annotation"),
                   remove_annotation_outliers = TRUE)
Hsap_gff_import <- read_gff(Hsap_gff)

## End(Not run)

```

---

getGFFSet

*GFF retrieval of multiple species*


---

## Description

Main GFF retrieval function for a set of organism of interest. By specifying the scientific names of the organisms of interest the corresponding fasta-files storing the GFF of the organisms of interest will be downloaded and stored locally. GFF files can be retrieved from several databases.

## Usage

```

getGFFSet(
  db = "refseq",
  organisms,
  reference = FALSE,
  release = NULL,
  skip_bacteria = TRUE,
  gunzip = TRUE,
  remove_annotation_outliers = FALSE,
  update = FALSE,
  format = "gff3",
  path = "set_GFF",
  mute_citation = FALSE
)

```

**Arguments**

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organisms	a character vector storing the names of the organisms than shall be retrieved. There are three available options to characterize an organism:
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they needs to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
gunzip	a logical, indicating whether or not files should be unzipped.
remove_annotation_outliers	shall outlier lines be removed from the input annotation_file? If yes, then the initial annotation_file will be overwritten and the removed outlier lines will be stored at <a href="#">tempdir</a> for further exploration.
update	logical, default FALSE. Updated backend cached files needed. Usually keep this false, to make ut run much faster. Only set to TRUE, if you believe you cache is outdated (Species only exist in newest release etc)
format	"gff3", alternative "gtf" for ensembl.
path	character, default location is paste0("set_", toupper(set_type))
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

**Details**

Internally this function loads the the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

and creates a directory 'set\_CDSs' to store the CDSs of interest as fasta files for future processing. In case the corresponding fasta file already exists within the 'set\_CDSs' folder and is accessible within the workspace, no download process will be performed.

**Value**

File path to downloaded genomes (names are identifiers: 'new' (file was downloaded now), 'old' files did already exist)



**Author(s)**

Hajk-Georg Drost

**See Also**

Other getBioSet: [getBioSet\(\)](#), [getCDSSet\(\)](#), [getCollectionSet\(\)](#), [getGenomeSet\(\)](#), [getProteomeSet\(\)](#), [getRNASET\(\)](#)

Other gff: [getGFF\(\)](#), [read\\_gff\(\)](#)

**Examples**

```
## Not run:
getBioSet("refseq", organisms = c("Arabidopsis thaliana",
                                "Arabidopsis lyrata",
                                "Capsella rubella"),
          set_type = "cds")

## End(Not run)
```

---

getGO

*Gene Ontology Query*

---

**Description**

This function takes a gene id as character vector from a given query organism and returns the corresponding GO terms and additional GO information.

**Usage**

```
getGO(organism, genes, filters, ...)
```

**Arguments**

organism	a character string specifying the scientific name of a query organism.
genes	a character vector storing the gene ids of a organisms of interest to be queried against Ensembl Biomart.
filters	a character vector specifying the filter (query key) for the Ensembl Biomart query, e.g. filter = "ensembl_gene_id".
...	additional parameters that can be passed to the <a href="#">biomart</a> function.

**Details**

This function takes the scientific name of a query organism, a set of genes for which GO terms and additional information shall be retrieved, and a filter argument that specifies the attribute for the query genes.

**Author(s)**

Hajk-Georg Drost

**See Also**

[biomart](#), [organismFilters](#), [organismBM](#), [getBM](#), [getMarts](#), [getDatasets](#), [getFilters](#)

**Examples**

```
## Not run:
GO_tbl <- getGO(organism = "Arabidopsis thaliana",
               genes     = c("AT1G06090", "AT1G06100"),
               filters   = "ensembl_gene_id")

# look at the result
head(GO_tbl)

## End(Not run)
```

---

getGroups	<i>Retrieve available groups for a kingdom of life (only available for NCBI RefSeq and NCBI Genbank)</i>
-----------	--

---

**Description**

A short list of available groups for a kingdom of life.

**Usage**

```
getGroups(db = "refseq", kingdom)
```

**Arguments**

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"><li>• db = "refseq"</li><li>• db = "genbank"</li></ul> Default is db = "refseq".
kingdom	a character string specifying for which kingdom of life groups shall be retrieved. See <a href="#">getKingdoms</a> for details.

**Author(s)**

Hajk-Georg Drost

**See Also**

[meta.retrieval](#), [getGenome](#), [getProteome](#), [getCDS](#), [getKingdoms](#)

## Examples

```
# get possible kigdom names
getKingdoms(db = "refseq")
## Not run:
# retrieve subgroups for vertebrate_mammalian available from refseq
getGroups(db = "refseq", kingdom = "vertebrate_mammalian")

# get possible kigdom names
getKingdoms(db = "genbank")
# retrieve subgroups for vertebrate_mammalian available from genbank
getGroups(db = "genbank", kingdom = "vertebrate_mammalian")

## End(Not run)
```

---

getGTF

*Genome Annotation Retrieval (GTF)*


---

## Description

Main retrieval function for GTF files of an organism of interest. By specifying the scientific name of an organism of interest the corresponding GTF file storing the annotation for the organism of interest can be downloaded and stored locally. GTF files can be retrieved from several databases.

## Usage

```
getGTF(
  db = "ensembl",
  organism,
  remove_annotation_outliers = FALSE,
  path = file.path("ensembl", "annotation"),
  release = NULL,
  mute_citation = FALSE
)
```

## Arguments

- |          |  |
|----------|--|
| db       | a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>   |
| organism | Organism selector id, there are three options to characterize an organism: <ul style="list-style-type: none"> <li>• by scientific name: e.g. organism = "Homo sapiens"</li> <li>• by database specific accession identifier: e.g. organism = "GCF_000001405.37" (= NCBI RefSeq identifier for Homo sapiens)</li> </ul> |

- by taxonomic identifier from NCBI Taxonomy: e.g. organism = "9606" (= taxid of Homo sapiens)

remove_annotation_outliers	shall outlier lines be removed from the input annotation_file? If yes, then the initial annotation_file will be overwritten and the removed outlier lines will be stored at <a href="#">tempdir</a> for further exploration.
path	a character string specifying the location (a folder) in which the corresponding annotation file shall be stored. Default is path = file.path("_ncbi_downloads", "annotation").
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

### Details

Internally this function loads the the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

and creates a directory relative to file type, if you get fasta genomes it will be '\_ncbi\_downloads/genomes'.

In case the corresponding fasta file already exists within the '\_ncbi\_downloads/genomes' folder and is accessible within the workspace, no download process will be performed. For other file types the same rule applies.

### Value

File path to downloaded genome.

### Author(s)

Hajk-Georg Drost

### See Also

Other getBio: [getBio\(\)](#), [getCDS\(\)](#), [getCollection\(\)](#), [getGenome\(\)](#), [getProteome\(\)](#), [getRNA\(\)](#)

Other gff: [getGFFSet\(\)](#), [read\\_gff\(\)](#)

### Examples

```
## Not run:
# download the annotation of Homo sapiens from ensembl
# and store the corresponding genome file in 'ensembl/annotation'
getGTF(db          = "ensembl",
       organism    = "Homo sapiens",
       path        = file.path("ensembl", "annotation"))

getGTF(db          = "ensembl",
```

```
organism      = "Homo sapiens",
path          = file.path("ensembl", "annotation"),
assembly_type = "primary_assembly")

## End(Not run)
```

---

getKingdomAssemblySummary

*Retrieve and summarise the assembly\_summary.txt files from NCBI for all kingdoms*

---

## Description

Retrieval function of the assembly\_summary.txt file from NCBI for all kingdoms. The assembly\_summary.txt files store available species on NCBI.

## Usage

```
getKingdomAssemblySummary(
  db,
  skip_bacteria = TRUE,
  file = assemblies_info_path(db)
)
```

## Arguments

db	database name. E.g. refseq or genbank.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
file	path, local path to total summary file, default is in tmp folder.

## Author(s)

Hajk-Georg Drost

## See Also

[getSummaryFile](#), [getMetaGenomeSummary](#), [get.ensembl.info](#)

### Examples

```
## Not run:
# This example will run the default version of this function
# whereby information for Bacteria are not downloaded
test <- getKingdomAssemblySummary(db = "genbank", skip_bacteria = TRUE)
test
# Users can then retrieve information for Bacteria by skip_bacteria = FALSE
test2 <- getKingdomAssemblySummary(db = "genbank", skip_bacteria = FALSE)
test2

## End(Not run)
```

---

getKingdoms

*Retrieve available kingdoms of life*

---

### Description

A short list of available kingdoms of life

### Usage

```
getKingdoms(db = "refseq")
```

### Arguments

db a character string specifying the database from which the genome shall be retrieved: db = "refseq", db = "genbank", db = "ensembl", db = "ensemblgenomes". Default is db = "refseq".

### Author(s)

Hajk-Georg Drost

### See Also

[meta.retrieval](#), [getGenome](#), [getProteome](#), [getCDS](#), [getGroups](#)

### Examples

```
# retrieve kingdoms available from refseq
getKingdoms(db = "refseq")

# retrieve kingdoms available from genbank
getKingdoms(db = "genbank")
```

---

getMarts	<i>Retrieve information about available Ensembl Biomart databases</i>
----------	---

---

**Description**

This function queries the Ensembl Biomart API and returns a table storing information about all available Ensembl Biomart databases.

**Usage**

```
getMarts()
```

**Author(s)**

Hajk-Georg Drost

**See Also**

[getDatasets](#), [getAttributes](#), [getFilters](#), [organismBM](#), [organismFilters](#), [organismAttributes](#)

**Examples**

```
## Not run:  
# get a table of all available databases from Ensembl Biomart  
getMarts()  
  
## End(Not run)
```

---

getMetaGenomeAnnotations	<i>Retrieve annotation *.gff files for metagenomes from NCBI Genbank</i>
--------------------------	--

---

**Description**

Retrieve available annotation \*.gff files for metagenomes from NCBI Genbank. NCBI Genbank allows users to download entire metagenomes and their annotations of several metagenome projects. This function downloads available metagenomes that can then be downloaded via [getMetaGenomes](#).

**Usage**

```
getMetaGenomeAnnotations(  
  name,  
  path = file.path("_ncbi_downloads", "metagenome", "annotations")  
)
```

**Arguments**

name metagenome name retrieved by [listMetaGenomes](#).  
path a character string specifying the location (a folder) in which the corresponding metagenome annotations shall be stored. Default is path = file.path("\_ncbi\_downloads", "metagenome")

**Author(s)**

Hajk-Georg Drost

**See Also**

[getMetaGenomes](#), [listMetaGenomes](#), [getGFF](#)

**Examples**

```
## Not run:  
# Frist, retrieve a list of available metagenomes  
listMetaGenomes()  
  
# Now, retrieve the 'human gut metagenome'  
getMetaGenomeAnnotations(name = "human gut metagenome")  
  
## End(Not run)
```

---

getMetaGenomes	<i>Retrieve metagenomes from NCBI Genbank</i>
----------------	---

---

**Description**

Retrieve available metagenomes from NCBI Genbank. NCBI Genbank allows users to download entire metagenomes of several metagenome projects. This function downloads available metagenomes that can then be downloaded via [getMetaGenomes](#).

**Usage**

```
getMetaGenomes(name, path = file.path("_ncbi_downloads", "metagenome"))
```

**Arguments**

name metagenome name retrieved by [listMetaGenomes](#).  
path a character string specifying the location (a folder) in which the corresponding metagenome shall be stored. Default is path = file.path("\_ncbi\_downloads", "metagenome").

**Author(s)**

Hajk-Georg Drost



**See Also**

[getMetaGenomeAnnotations](#), [listMetaGenomes](#)

**Examples**

```
## Not run:  
# Frist, retrieve a list of available metagenomes  
listMetaGenomes()  
  
# Now, retrieve the 'human gut metagenome'  
getMetaGenomes(name = "human gut metagenome")  
  
## End(Not run)
```

---

getMetaGenomeSummary *Retrieve the assembly\_summary.txt file from NCBI genbank metagenomes*

---

**Description**

Retrieval function of the assembly\_summary.txt file from NCBI genbank metagenomes. This files stores all available metagenome projects on NCBI Genbank.

**Usage**

```
getMetaGenomeSummary(  
  local_file = file.path(cachedir(), "assembly_summary_metagenomes_genbank.txt")  
)
```

**Arguments**

local\_file      where to store this backend file, default: file.path(cachedir(), "assembly\_summary\_metagenomes\_genbank.txt")

**Author(s)**

Hajk-Georg Drost

**See Also**

[getKingdomAssemblySummary](#), [getSummaryFile](#)

**Examples**

```
## Not run:  
meta.summary <- getMetaGenomeSummary()  
meta.summary  
  
## End(Not run)
```

---

 getProteome

*Proteome Retrieval*


---

### Description

Main proteome retrieval function for an organism of interest. By specifying the scientific name of an organism of interest the corresponding fasta-file storing the proteome of the organism of interest can be downloaded and stored locally. Proteome files can be retrieved from several databases.

### Usage

```
getProteome(
  db = "refseq",
  organism,
  reference = TRUE,
  skip_bacteria = TRUE,
  release = NULL,
  gunzip = FALSE,
  update = TRUE,
  path = file.path("_ncbi_downloads", "proteomes"),
  mute_citation = FALSE
)
```

### Arguments

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organism	Organism selector id, there are three options to characterize an organism: <ul style="list-style-type: none"> <li>• by scientific name: e.g. organism = "Homo sapiens"</li> <li>• by database specific accession identifier: e.g. organism = "GCF_000001405.37" (= NCBI RefSeq identifier for Homo sapiens)</li> <li>• by taxonomic identifier from NCBI Taxonomy: e.g. organism = "9606" (= taxid of Homo sapiens)</li> </ul>
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.

release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
gunzip	a logical, indicating whether or not files should be unzipped.
update	logical, default TRUE. (Uniprot only for now!) If species info file exists already, do not re download, makes it faster but the file can be old, i.e. no longer as complete as it could be.
path	a character string specifying the location (a folder) in which the corresponding proteome shall be stored. Default is path = file.path("_ncbi_downloads", "proteomes").
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

### Details

Internally this function loads the the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

and creates a directory '\_ncbi\_downloads/proteomes' to store the proteome of interest as fasta file for future processing.

### Value

File path to downloaded proteome.

### Author(s)

Hajk-Georg Drost

### See Also

Other getBio: [getBio\(\)](#), [getCDS\(\)](#), [getCollection\(\)](#), [getGFF\(\)](#), [getGenome\(\)](#), [getRNA\(\)](#)

Other proteome: [getProteomeSet\(\)](#), [read\\_proteome\(\)](#)

### Examples

```
## Not run:
# download the proteome of Arabidopsis thaliana from NCBI RefSeq
# and store the corresponding proteome file in '_ncbi_downloads/refseq/proteomes'
file_path <- getProteome( db      = "refseq",
                        organism = "Arabidopsis thaliana",
                        path     = file.path("_ncbi_downloads", "refseq", "proteomes") )
# import proteome into R session
Ath_proteome <- read_proteome(file_path, format = "fasta")

# download the proteome of Arabidopsis thaliana from NCBI Genbank
# and store the corresponding proteome file in '_ncbi_downloads/genbank/proteomes'
file_path <- getProteome( db      = "genbank",
```

```

        organism = "Arabidopsis thaliana",
        path      = file.path("_ncbi_downloads", "genbank", "proteomes") )
# import proteome into R session
Ath_proteome <- read_proteome(file_path, format = "fasta")

# and store the corresponding proteome file in '_downloads/uniprot/proteomes'
file_path <- getProteome( db      = "uniprot",
                        organism = "Arabidopsis thaliana",
                        path      = file.path("_downloads", "uniprot", "proteomes") )
# import proteome into R session
Ath_proteome <- read_proteome(file_path, format = "fasta")

# download the proteome of Arabidopsis thaliana from ENSEMBL
# and store the corresponding proteome file in '_downloads/ensembl/proteomes'
file_path <- getProteome( db      = "ensembl",
                        organism = "Arabidopsis thaliana",
                        path      = file.path("_downloads", "ensembl", "proteomes") )
# import proteome into R session
Ath_proteome <- read_proteome(file_path, format = "fasta")

## End(Not run)

```

---

getProteomeSet

*Proteome retrieval of multiple species*


---

## Description

Main proteome retrieval function for a set of organism of interest. By specifying the scientific names of the organisms of interest the corresponding fasta-files storing the proteome of the organisms of interest will be downloaded and stored locally. proteome files can be retrieved from several databases.

## Usage

```

getProteomeSet(
  db = "refseq",
  organisms,
  reference = FALSE,
  release = NULL,
  skip_bacteria = TRUE,
  gunzip = TRUE,
  update = FALSE,
  path = "set_proteomes",
  mute_citation = FALSE
)

```

**Arguments**

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organisms	a character vector storing the names of the organisms than shall be retrieved. There are three available options to characterize an organism:
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they needs to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
gunzip	a logical, indicating whether or not files should be unzipped.
update	logical, default FALSE. Updated backend cached files needed. Usually keep this false, to make ut run much faster. Only set to TRUE, if you believe you cache is outdated (Species only exist in newest release etc)
path	a character string specifying the location (a folder) in which the corresponding proteomes shall be stored. Default is path = "set_proteomes".
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

**Details**

Internally this function loads the the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

and creates a directory 'set\_CDSs' to store the CDSs of interest as fasta files for future processing. In case the corresponding fasta file already exists within the 'set\_CDSs' folder and is accessible within the workspace, no download process will be performed.

**Value**

File path to downloaded genomes (names are identifiers: 'new' (file was downloaded now), 'old' files did already exist)

**Author(s)**

Hajk-Georg Drost

**See Also**

Other getBioSet: [getBioSet\(\)](#), [getCDSSet\(\)](#), [getCollectionSet\(\)](#), [getGFFSet\(\)](#), [getGenomeSet\(\)](#), [getRNASet\(\)](#)

Other proteome: [getProteome\(\)](#), [read\\_proteome\(\)](#)

**Examples**

```
## Not run:
# download the proteomes of three different species at the same time
#### Database: NCBI RefSeq
file_paths <- getProteomeSet(db = "refseq", organisms = c("Arabidopsis thaliana",
                                                       "Arabidopsis lyrata",
                                                       "Capsella rubella"))

# look at file paths
file_paths

#### Database: NCBI Genbank
file_paths <- getProteomeSet(db = "genbank", organisms = c("Arabidopsis thaliana",
                                                         "Arabidopsis lyrata",
                                                         "Capsella rubella"))

# look at file paths
file_paths

# download the proteomes of three different species at the same time
#### Database: ENSEMBL
file_paths <- getProteomeSet(db = "ensembl", organisms = c("Homo sapiens",
                                                         "Mus musculus",
                                                         "Caenorhabditis elegans"))

# look at file paths
file_paths

# download the proteomes of three different species at the same time
#### Database: UniProt
file_paths <- getProteomeSet(db = "uniprot", organisms = c("Homo sapiens",
                                                         "Mus musculus",
                                                         "Caenorhabditis elegans"))

# look at file paths
file_paths

## End(Not run)
```

---

getReleases

*Retrieve available database releases or versions of ENSEMBL*

---

**Description**

Retrieve available database releases or versions of ENSEMBL.

**Usage**

```
getReleases(db = "ensembl")
```

**Arguments**

db a character string specifying the database from which available release versions shall be retrieved:

- db = "ensembl"

**Author(s)**

Hajk-Georg Drost

**Examples**

```
## Not run:  
# retrieve available release versions of ENSEMBL  
getReleases("ensembl")  
  
## End(Not run)
```

---

getRepeatMasker	<i>Repeat Masker Retrieval</i>
-----------------	--------------------------------

---

**Description**

Main Repeat Masker output retrieval function for an organism of interest. By specifying the scientific name of an organism of interest the corresponding Repeat Masker file storing the genome of the organism of interest can be downloaded and stored locally. Repeat Masker files can be retrieved from several databases.

**Usage**

```
getRepeatMasker(  
  db = "refseq",  
  organism,  
  reference = FALSE,  
  skip_bacteria = TRUE,  
  release = NULL,  
  gunzip = FALSE,  
  path = file.path("_ncbi_downloads", "repeatmasker"),  
  mute_citation = FALSE  
)
```

**Arguments**

<code>db</code>	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• <code>db = "refseq"</code></li> <li>• <code>db = "genbank"</code></li> </ul>
<code>organism</code>	a character string specifying the scientific name of the organism of interest, e.g. <code>organism = "Homo sapiens"</code> .
<code>reference</code>	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
<code>skip_bacteria</code>	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify <code>skip_bacteria = FALSE</code> . When <code>skip_bacteria = FALSE</code> is set then the bacterial summary file will be downloaded.
<code>release</code>	most recent database version is used. <code>release = 75</code> would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data in the standard format before that.
<code>gunzip</code>	a logical, indicating whether or not files should be unzipped.
<code>path</code>	a character string specifying the location (a folder) in which the corresponding file shall be stored. Default is <code>path = file.path("_ncbi_downloads", "repeatmasker")</code> .
<code>mute_citation</code>	logical value indicating whether citation message should be muted.

**Details**

Internally this function loads the `overview.txt` file from NCBI:

`refseq: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/`

`genbank: ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/`

and creates a directory `'_ncbi_downloads/repeatmasker'` to store the files of interest as fasta file for future processing. In case the corresponding fasta file already exists within the `'_ncbi_downloads/repeatmasker'` folder and is accessible within the workspace, no download process will be performed.

**Value**

File path to downloaded Repeat Masker output file.

**Author(s)**

Hajk-Georg Drost

**See Also**

[getGenome](#), [getProteome](#), [getCDS](#), [getGFF](#), [getRNA](#), [getCollection](#), [meta.retrieval](#), [read\\_rm](#)



## Examples

```
## Not run:

# download the Repeat Masker output file of Homo sapiens from refseq
# and store the corresponding genome file in '_ncbi_downloads/genomes'
file_path <- getRepeatMasker( db      = "refseq",
                             organism = "Homo sapiens",
                             path    = file.path("_ncbi_downloads", "repeatmasker"))

Hsap_repeatmasker <- read_rm(file_path)

## End(Not run)
```

---

getRNA

*RNA Sequence Retrieval*

---

## Description

Main retrieval function for RNA sequences of an organism of interest. By specifying the scientific name of an organism of interest the corresponding fasta-file storing the RNA information for the organism of interest can be downloaded and stored locally. RNA files can be retrieved from several databases.

## Usage

```
getRNA(
  db = "refseq",
  organism,
  reference = FALSE,
  skip_bacteria = TRUE,
  release = NULL,
  assembly_type = "toplevel",
  path = file.path("_ncbi_downloads", "RNA"),
  gunzip = FALSE,
  mute_citation = FALSE
)
```

## Arguments

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"><li>• db = "refseq"</li><li>• db = "genbank"</li><li>• db = "ensembl"</li></ul>
organism	Organism selector id, there are three options to characterize an organism:

	<ul style="list-style-type: none"> <li>• by scientific name: e.g. <code>organism = "Homo sapiens"</code></li> <li>• by database specific accession identifier: e.g. <code>organism = "GCF_000001405.37"</code> (= NCBI RefSeq identifier for Homo sapiens)</li> <li>• by taxonomic identifier from NCBI Taxonomy: e.g. <code>organism = "9606"</code> (= taxid of Homo sapiens)</li> </ul>
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify <code>skip_bacteria = FALSE</code> . When <code>skip_bacteria = FALSE</code> is set then the bacterial summary file will be downloaded.
release	a numeric, the database release version of ENSEMBL ( <code>db = "ensembl"</code> ). Default is <code>release = NULL</code> meaning that the most recent database version is used. <code>release = 75</code> would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
assembly_type	a character, default "toplevel". id type of assembly, either "toplevel" or "primary_assembly" usually.
path	a character string specifying the location (a folder) in which the corresponding CDS file shall be stored. Default is <code>path = file.path("_ncbi_downloads", "RNA")</code> .
gunzip	a logical, indicating whether or not files should be unzipped.
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

### Details

Internally this function loads the the overview.txt file from NCBI:

refseq: `ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/`

genbank: `ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/`

and creates a directory relative to file type, if you get fasta genomes it will be `_ncbi_downloads/genomes'`.

In case the corresponding fasta file already exists within the `'_ncbi_downloads/genomes'` folder and is accessible within the workspace, no download process will be performed. For other file types the same rule applies.

### Value

File path to downloaded genome.

### Author(s)

Hajk-Georg Drost

### See Also

Other getBio: `getBio()`, `getCDS()`, `getCollection()`, `getGFF()`, `getGenome()`, `getProteome()`

Other rna: `getRNASet()`, `read_rna()`

## Examples

```
## Not run:
# download the RNA of Arabidopsis thaliana from refseq
# and store the corresponding RNA file in '_ncbi_downloads/RNA'
file_path <- getRNA( db      = "refseq",
                    organism = "Arabidopsis thaliana",
                    path    = file.path("_ncbi_downloads", "RNA"))

Ath_RNA <- read_rna(file_path, format = "fasta")

## End(Not run)
```

---

getRNASet

*RNA Retrieval of multiple species*

---

## Description

Main RNA retrieval function for a set of organism of interest. By specifying the scientific names of the organisms of interest the corresponding fasta-files storing the RNA of the organisms of interest will be downloaded and stored locally. RNA files can be retrieved from several databases.

## Usage

```
getRNASet(
  db = "refseq",
  organisms,
  reference = FALSE,
  release = NULL,
  skip_bacteria = TRUE,
  gunzip = TRUE,
  update = FALSE,
  path = "set_RNAs",
  mute_citation = FALSE
)
```

## Arguments

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
organisms	a character vector storing the names of the organisms than shall be retrieved. There are three available options to characterize an organism:
reference	a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome.

release	a numeric, the database release version of ENSEMBL (db = "ensembl"). Default is release = NULL meaning that the most recent database version is used. release = 75 would for human would give the stable GRCh37 release in ensembl. Value must be > 46, since ensembl did not structure their data if the standard format before that.
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they needs to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
gunzip	a logical, indicating whether or not files should be unzipped.
update	logical, default FALSE. Updated backend cached files needed. Usually keep this false, to make ut run much faster. Only set to TRUE, if you believe you cache is outdated (Species only exist in newest release etc)
path	a character string specifying the location (a folder) in which the corresponding RNAs shall be stored. Default is path = "set_RNAs".
mute_citation	logical, default FALSE, indicating whether citation message should be muted.

### Details

Internally this function loads the the overview.txt file from NCBI:

refseq: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

genbank: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>

and creates a directory 'set\_CDSs' to store the CDSs of interest as fasta files for future processing. In case the corresponding fasta file already exists within the 'set\_CDSs' folder and is accessible within the workspace, no download process will be performed.

### Value

File path to downloaded genomes (names are identifiers: 'new' (file was downloaded now), 'old' files did already exist)

### Author(s)

Hajk-Georg Drost

### See Also

Other getBioSet: [getBioSet\(\)](#), [getCDSSet\(\)](#), [getCollectionSet\(\)](#), [getGFFSet\(\)](#), [getGenomeSet\(\)](#), [getProteomeSet\(\)](#)

Other rna: [getRNA\(\)](#), [read\\_rna\(\)](#)

## Examples

```
## Not run:
getBioSet("refseq", organisms = c("Arabidopsis thaliana",
                                  "Arabidopsis lyrata",
                                  "Capsella rubella"),
          set_type = "cds")

## End(Not run)
```

---

getSummaryFile	<i>Helper function to retrieve the assembly_summary.txt file from NCBI</i>
----------------	--

---

## Description

Retrieval function of the assembly\_summary.txt file from NCBI.

## Usage

```
getSummaryFile(db, kingdom, file = assemblies_info_path(db, kingdom))
```

## Arguments

db	database name. E.g. refseq or genbank.
kingdom	kingdom for which assembly_summary.txt file shall be retrieved. See also <a href="#">getKingdoms</a> .
file	path, local path to total summary file, default is in tmp folder.

## Author(s)

Hajk-Georg Drost

## See Also

[getKingdomAssemblySummary](#), [getMetaGenomeSummary](#)

## Examples

```
## Not run:
test <- getSummaryFile("refseq", "plant")
test

## End(Not run)
```

---

getUniProtInfo	<i>Get uniprot info from organism</i>
----------------	---------------------------------------

---

**Description**

Get uniprot info from organism

**Usage**

```
getUniProtInfo(organism, path = cachedir(), update = TRUE)
```

**Arguments**

organism	character, name of organism
path	path at which the info file shall be stored locally.
update	shall the internal <code>cachedir</code> file be deleted and the info file freshly downloaded from the UniProt API?

---

getUniProtSTATS	<i>Retrieve UniProt Database Information File (STATS)</i>
-----------------	---

---

**Description**

The UniProt stores a STATS file to summarise all available information for their reference proteomes. Users can now download this file and process it with `biomartr`.

**Usage**

```
getUniProtSTATS(update = FALSE)
```

**Arguments**

update	shall the internal <code>cachedir</code> file be deleted and the STATS file freshly downloaded from the UniProt FTP servers?
--------	--

**Author(s)**

Hajk-Georg Drost

**Examples**

```
## Not run:  
# retrieve STATS file from UniProt  
uniprot_info <- getUniProtSTATS(update = TRUE)  
# look at results  
uniprot_info  
  
## End(Not run)
```

---

is.genome.available     *Check Genome Availability*

---

### Description

This function checks the availability of a given genome on the NCBI servers specified as scientific name.

### Usage

```
is.genome.available(  
  db = "refseq",  
  organism,  
  skip_bacteria = TRUE,  
  details = FALSE  
)
```

### Arguments

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"><li>• db = "refseq"</li><li>• db = "genbank"</li><li>• db = "ensembl"</li><li>• db = "uniprot"</li></ul>
organism	there are three options to characterize an organism: <ul style="list-style-type: none"><li>• by scientific name: e.g. organism = "Homo sapiens"</li><li>• by database specific accession identifier: e.g. organism = "GCF_000001405.37" (= NCBI RefSeq identifier for Homo sapiens)</li><li>• by taxonomic identifier from NCBI Taxonomy: e.g. organism = "9606" (= taxid of Homo sapiens)</li></ul>
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.
details	a logical value specifying whether or not details on genome size, kingdom, etc. shall be printed to the console instead of a boolean value.

### Details

Internally this function calls the [listGenomes](#) function to detect all available genomes and checks whether or not the specified organism is available for download.

**Value**

a logical value specifying whether or not the genome of the input organism is available. In case `details = TRUE` only a character string specifying the genome details is being returned.

**Author(s)**

Hajk-Georg Drost

**Examples**

```
## Not run:
# checking whether the Homo sapiens genome is stored on NCBI
is.genome.available(organism = "Homo sapiens", db = "refseq")

# and printing details
is.genome.available(organism = "Homo sapiens", db = "refseq", details = TRUE)

# checking whether the Homo sapiens genome is stored on ENSEMBL
is.genome.available(organism = "Homo sapiens", db = "ensembl")

# and printing details
is.genome.available(organism = "Homo sapiens",
                    details = TRUE,
                    db = "ensembl")

## End(Not run)
```

---

listDatabases

*Retrieve a List of Available NCBI Databases for Download*

---

**Description**

This function allows you to retrieve a list of database names and versions that can be downloaded from corresponding servers.

Database retrieval is crucial for most biological studies and analyses. There is a vast diversity of databases that can be accessed remotely or that can be downloaded to your local machine. This function provides an interface to databases that can be downloaded from NCBI servers and lists all available databases and their database version to be able to select an appropriate database for download with [download.database](#).

**Usage**

```
listDatabases(db = "nr", update = FALSE)

listNCBIDatabases(db = "nr", update = FALSE)
```



**Arguments**

db a character string specifying the name of the database that shall be searched for.  
update a logical value specifying whether or not the local listDatabases.txt file shall be updated by remote access to NCBI.

**Author(s)**

Hajk-Georg Drost

**See Also**

[download.database](#), [download.database.all](#)

**Examples**

```
## Not run:  
# retrieve all versions of the NCBI 'nr' database that can be downloaded  
listNCBIDatabases(db = "nr")  
  
# analogous:  
# listNCBIDatabases(db = "cdd")  
# listNCBIDatabases(db = "nt")  
# listNCBIDatabases(db = "gss")  
# listNCBIDatabases(db = "refseq_protein")  
  
## End(Not run)
```

---

listGenomes

*List All Available Genomes either by kingdom, group, or subgroup*

---

**Description**

This function retrieves the names of all genomes available on the NCBI ftp:// server and stores the results in a file named 'overview.txt' inside the directory '\_ncbi\_downloads' that is built inside the workspace.

**Usage**

```
listGenomes(  
  db = "refseq",  
  type = "all",  
  subset = NULL,  
  details = FALSE,  
  skip_bacteria = FALSE  
)
```

**Arguments**

db	a character string specifying the database for which genome availability shall be checked. Available options are: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> </ul>
type	a character string specifying a potential filter of available genomes. Available options are: <ul style="list-style-type: none"> <li>• type = "all"</li> <li>• type = "kingdom"</li> <li>• type = "group"</li> <li>• type = "subgroup"</li> </ul>
subset	a character string or character vector specifying a subset of type. E.g. if users are interested in retrieving all Eukaryota species, they can specify: type = "kingdom" and subset = "Eukaryota".
details	a boolean value specifying whether only the scientific names of stored genomes shall be returned (details = FALSE) or all information such as <ul style="list-style-type: none"> <li>• organism_name</li> <li>• kingdoms</li> <li>• group</li> <li>• subgroup</li> <li>• file_size_MB, etc.</li> </ul>
skip_bacteria	Due to its enormous dataset size (> 700MB as of July 2023), the bacterial summary file will not be loaded by default anymore. If users wish to gain insights for the bacterial kingdom they need to actively specify skip_bacteria = FALSE. When skip_bacteria = FALSE is set then the bacterial summary file will be downloaded.

**Details**

Internally this function loads the overview.txt file from NCBI and creates a directory '\_ncbi\_downloads' in the tempdir() folder to store the overview.txt file for future processing. In case the overview.txt file already exists within the '\_ncbi\_downloads' folder and is accessible within the workspace, no download process will be performed again.

**Note**

Please note that the ftp:// connection relies on the NCBI or ENSEMBL server and cannot be accurately accessed via a proxy.

**Author(s)**

Hajk-Georg Drost

**Examples**

```
## Not run:
# print details for refseq
listGenomes(db = "refseq")
# print details for all plants in refseq
listGenomes(db = "refseq", type = "kingdom")
# print details for all plant groups in refseq
listGenomes(db = "refseq", type = "group")
# print details for all plant subgroups in refseq
listGenomes(db = "refseq", type = "subgroup")
# Ensembl
listGenomes(db = "ensembl", type = "kingdom", subset = "EnsemblVertebrates")

## End(Not run)
```

---

**listGroups***List number of available genomes in each taxonomic group*

---

**Description**

Users can retrieve the available number of sequenced genomes per group. Only available for db = "refseq" and db = "genbank".

**Usage**

```
listGroups(db = "refseq", kingdom = "all", details = FALSE)
```

**Arguments**

db	a character string specifying the database for which genome availability shall be checked. Available options are: <ul style="list-style-type: none"><li>• db = "refseq"</li><li>• db = "genbank"</li></ul>
kingdom	a kingdom specification retrieved by <a href="#">getKingdoms</a> .
details	shall all species corresponding to the specified kingdom be returned? Default is details = FALSE.

**Author(s)**

Hajk-Georg Drost

**See Also**

[listGenomes](#), [is.genome.available](#), [listKingdoms](#)

**Examples**

```
## Not run:
# example for refseq
listGroups(db = "refseq")
# example for genbank
listGroups(db = "genbank")
### in case groups should be specified by kingdom
# first, retrieve available kingdom names
listKingdoms()
# now we choose kingdom "bacteria"
listGroups(db = "refseq", kingdom = "bacteria")
# or
listGroups(db = "genbank", kingdom = "bacteria")

## End(Not run)
```

---

listKingdoms

*List number of available genomes in each kingdom of life*


---

**Description**

Users can retrieve the available number of sequenced genomes per kingdom.

**Usage**

```
listKingdoms(db = "refseq")
```

**Arguments**

db a character string specifying the database for which genome availability shall be checked, e.g. db = "refseq", db = "genbank", db = "ensembl".

**Author(s)**

Hajk-Georg Drost

**See Also**

[listGenomes](#), [is.genome.available](#), [listGroups](#)

**Examples**

```
## Not run:
# list number of available genomes in refseq for each kingdom of life
listKingdoms(db = "refseq")
# example for genbank
listKingdoms(db = "genbank")
# example for ensembl
listKingdoms(db = "ensembl")
```

```
## End(Not run)
```

---

listMetaGenomes	<i>List available metagenomes on NCBI Genbank</i>
-----------------	---

---

### Description

List available metagenomes on NCBI genbank. NCBI genbank allows users to download entire metagenomes of several metagenome projects. This function lists all available metagenomes that can then be downloaded via [getMetaGenomes](#).

### Usage

```
listMetaGenomes(details = FALSE)
```

### Arguments

details	a boolean value specifying whether only the scientific names of stored metagenomes shall be returned (details = FALSE) or all information such as "organism_name", "bioproject", etc (details = TRUE).
---------	--

### Author(s)

Hajk-Georg Drost

### See Also

[getMetaGenomes](#), [getMetaGenomeSummary](#)

### Examples

```
## Not run:  
# retrieve available metagenome projects at NCBI Genbank  
listMetaGenomes()  
# retrieve detailed information on available metagenome projects  
# at NCBI Genbank  
listMetaGenomes(details = TRUE)  
  
## End(Not run)
```

---

meta.retrieval	<i>Perform Meta-Genome Retrieval</i>
----------------	--------------------------------------

---

### Description

Download genomes, proteomes, cds, gff, rna, or assembly stats files of all species within a kingdom of life.

### Usage

```
meta.retrieval(  
  db = "refseq",  
  kingdom,  
  group = NULL,  
  type = "genome",  
  restart_at_last = TRUE,  
  reference = FALSE,  
  combine = FALSE,  
  path = NULL  
)
```

### Arguments

db	a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"><li>• db = "refseq"</li><li>• db = "genbank"</li><li>• db = "emsembl"</li></ul>
kingdom	a character string specifying the kingdom of the organisms of interest, e.g. <ul style="list-style-type: none"><li>• For NCBI RefSeq:<ul style="list-style-type: none"><li>– kingdom = "archaea"</li><li>– kingdom = "bacteria"</li><li>– kingdom = "fungi"</li><li>– kingdom = "invertebrate"</li><li>– kingdom = "plant"</li><li>– kingdom = "protozoa"</li><li>– kingdom = "viral"</li><li>– kingdom = "vertebrate_mammalian"</li><li>– kingdom = "vertebrate_other"</li></ul></li><li>• For NCBI Genbank:<ul style="list-style-type: none"><li>– kingdom = "archaea"</li><li>– kingdom = "bacteria"</li><li>– kingdom = "fungi"</li></ul></li></ul>

- kingdom = "invertebrate"
- kingdom = "plant"
- kingdom = "protozoa"
- kingdom = "vertebrate\_mammalian"
- kingdom = "vertebrate\_other"
- For ENSEMBL:
  - kingdom = "EnsemblVertebrates"
  - kingdom = "EnsemblPlants"
  - kingdom = "EnsemblFungi"
  - kingdom = "EnsemblMetazoa"
  - kingdom = "EnsemblBacteria"
  - kingdom = "EnsemblProtists"

Available kingdoms can be retrieved with [getKingdoms](#).

group	only species belonging to this subgroup will be downloaded. Groups can be retrieved with <a href="#">getGroups</a> .
type	<p>type of sequences that shall be retrieved. Options are:</p> <ul style="list-style-type: none"> <li>• type = "genome" : (for genome assembly retrieval; see also <a href="#">getGenome</a>),</li> <li>• type = "proteome" : (for proteome retrieval; see also <a href="#">getProteome</a>),</li> <li>• type = "cds" : (for coding sequence retrieval; see also <a href="#">getCDS</a>),</li> <li>• type = "gff" : (for annotation file retrieval in gff format; see also <a href="#">getGFF</a>),</li> <li>• type = "gtf" : (for annotation file retrieval in gtf format (only for ensembl and ensemblgenomes); see also <a href="#">getGTF</a>)</li> <li>• type = "rna" : (for RNA file retrieval in fasta format; see also <a href="#">getRNA</a>),</li> <li>• type = "rm" : (for Repeat Masker output file retrieval; see also <a href="#">getRepeatMasker</a>),</li> <li>• type = "assemblystats" : (for genome assembly quality stats file retrieval; see also <a href="#">getAssemblyStats</a>).</li> </ul>
restart_at_last	<p>a logical value indicating whether or not <code>meta.retrieval</code> should pick up at the last species when re-running the function.</p> <ul style="list-style-type: none"> <li>• If <code>restart_at_last = TRUE</code> (Default) then <code>meta.retrieval</code> will skip all organisms that are already present in the folder and will start downloading all remaining species. However, this way <code>meta.wretrieval</code> will not be able to check whether already downloaded organism files are corrupted or not by checking the md5 checksum.</li> <li>• If <code>restart_at_last = FALSE</code> then <code>meta.retrieval</code> will start from the beginning and crawl through already downloaded organism files and check whether already downloaded organism files are corrupted or not by checking the md5 checksum. After checking existing files the function will start downloading all remaining organisms.</li> </ul>
reference	<p>a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome. Options are:</p> <ul style="list-style-type: none"> <li>• <code>reference = FALSE</code> (Default): all organisms (reference, representative, and non-representative genomes) are downloaded.</li> </ul>

- `reference = TRUE`: organisms that are downloaded must be either a reference or representative genome. Thus, most genomes which are usually non-reference genomes will not be downloaded.
- `combine` just in case `type = "assemblystats"` is specified, shall assembly stats of individual species be imported and combined to a [data.frame](#)?
- `path` path to the folder in which downloaded genomes shall be stored. By default the kingdom name is used to name the output folder.

### Details

This function aims to perform bulk retrieval of the genomes, proteomes, cds, etc. of species that belong to the same kingdom of life or to the same subgroup.

### Value

a character vector storing the file paths of the retrieved files.

### Author(s)

Hajk-Georg Drost

### See Also

Other meta\_retrival: [meta.retrieval.all\(\)](#)

### Examples

```
## Not run:
# get all available kingdoms for refseq
getKingdoms(db = "refseq")
# download all vertebrate genomes from refseq
meta.retrieval(kingdom = "vertebrate_mammalian",
               db = "refseq",
               type = "genome")

# get all available kingdoms for genbank
getKingdoms(db = "genbank")
# download all vertebrate genomes from genbank
meta.retrieval(kingdom = "vertebrate_mammalian",
               db = "genbank",
               type = "genome")

# In case users do not wish to retrieve genomes from an entire kingdom,
# but rather from a subgroup (e.g. from species belonging to the
# Gammaproteobacteria class, a subgroup of the bacteria kingdom),
# they can use the following workflow"
# First, users can again consult the getKingdoms() function to retrieve
# kingdom information.
getKingdoms(db = "refseq")
```



```

# In this example, we will choose the bacteria kingdom.
# Now, the getGroups() function allows users to obtain available
# subgroups of the bacteria kingdom.
getGroups(db = "refseq", kingdom = "bacteria")

# Now we choose the group Gammaproteobacteria and specify
# the group argument in the meta.retrieval() function
meta.retrieval(kingdom = "bacteria",
               roup = "Gammaproteobacteria",
               db = "refseq",
               type = "genome")

## End(Not run)

```

---

meta.retrieval.all	<i>Perform Meta-Genome Retrieval of all organisms in all kingdoms of life</i>
--------------------	---

---

## Description

Download genomes, proteomes, cds, gff, rna, or assembly stats files of individual species of all kingdoms of life.

## Usage

```
meta.retrieval.all(db = "refseq", type = "genome", reference = FALSE)
```

## Arguments

- |      |  |
|------|--|
| db   | a character string specifying the database from which the genome shall be retrieved: <ul style="list-style-type: none"> <li>• db = "refseq"</li> <li>• db = "genbank"</li> <li>• db = "ensembl"</li> <li>• db = "ensemblgenomes"</li> </ul>  |
| type | type of sequences that shall be retrieved. Options are: <ul style="list-style-type: none"> <li>• type = "genome" : for genome assembly retrieval; see also <a href="#">getGenome</a>),</li> <li>• type = "proteome" : (for proteome retrieval; see also <a href="#">getProteome</a>),</li> <li>• type = "cds" : (for coding sequence retrieval; see also <a href="#">getCDS</a>),</li> <li>• type = "gff" : (for annotation file retrieval in gff format; see also <a href="#">getGFF</a>),</li> <li>• type = "gtf" : (for annotation file retrieval in gtf format (only for ensembl and ensemblgenomes); see also <a href="#">getGTF</a>),</li> <li>• type = "rna" : (for RNA file retrieval in fasta format; see also <a href="#">getRNA</a>),</li> <li>• type = "rm" : (for Repeat Masker output file retrieval; see also <a href="#">getRepeatMasker</a>),</li> <li>• type = "assemblystats" (for genome assembly quality stats file retrieval; see also <a href="#">getAssemblyStats</a>).</li> </ul> |

- reference a logical value indicating whether or not a genome shall be downloaded if it isn't marked in the database as either a reference genome or a representative genome. Options are:
- reference = FALSE (Default): all organisms (reference, representative, and non-representative genomes) are downloaded.
  - reference = TRUE: organisms that are downloaded must be either a reference or representative genome. Thus, most genomes which are usually non-reference genomes will not be downloaded.

### Details

This function aims to perform bulk retrieval of all genomes of species for all kingdoms of life.

### Value

a character vector storing the file paths of the retrieved files.

### Author(s)

Hajk-Georg Drost

### See Also

Other meta\_retrival: [meta.retrieval\(\)](#)

### Examples

```
## Not run:  
# download all genomes from refseq  
meta.retrieval.all(db = "refseq", type = "genome")  
# download all vertebrate genomes from genbank  
meta.retrieval.all(db = "genbank", type = "genome")  
# download all vertebrate genomes from ensemblgenomes  
meta.retrieval.all(db = "genbank", type = "ensemblgenomes")  
  
## End(Not run)
```

---

organismAttributes      *Retrieve Ensembl Biomart attributes for a query organism*

---

### Description

In addition to the [organismBM](#) function, this function returns all available attributes that can be accessed through different marts and datasets for a given query organism.

### Usage

```
organismAttributes(organism, update = FALSE, topic = NULL)
```

**Arguments**

organism	a character string specifying the scientific name of a query organism.
update	a logical value specifying whether or not the local listMart.txt, listDatasets.txt, and listAttributes_organism.txt files shall be updated by remote access to BioMart.
topic	a character string specifying a topic (category) of attributes, e.g. topic = "id".

**Details**

For a given query organism, this function retrieves all available attributes that can be accessed through different marts and datasets.

Sometimes the same attribute names correspond to different datasets and marts causing problems when using [getMarts](#). The approach introduced by this function provides (again) a organism centric way of accessing organism specific attributes.

The topic argument allows the user to search for specific attribute topics/categories for faster filtering.

**Value**

a data.frame storing corresponding attribute names, description, datasets, and marts.

**Note**

When you run this function for the first time, the data retrieval procedure will take some time, due to the remote access to BioMart. The corresponding result is then saved in a \*.txt file within the [tempdir](#) directory named "\_biomart/listMarts.txt", "\_biomart/listDatasets.txt", and "\_biomart/listAttributes\_organism.txt", allowing subsequent queries to perform much faster.

**Author(s)**

Hajk-Georg Drost

**References**

<http://biomart.org/>

Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman, Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).

BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma and Wolfgang Huber, Bioinformatics 21, 3439-3440 (2005).

**See Also**

[organismFilters](#), [organismBM](#), [biomart](#), [listAttributes](#)

### Examples

```
## Not run:  
# search for attribute topic id  
head(organismAttributes("Homo sapiens", topic = "id"), 20)  
  
## End(Not run)
```

---

organismBM

*Retrieve Ensembl Biomart marts and datasets for a query organism*

---

### Description

This function returns either all available biomart connections for all available organisms for which biomart access is possible, or (when specified) returns all organism specific biomart connections.

### Usage

```
organismBM(organism = NULL, update = FALSE, mute_citation = TRUE)
```

### Arguments

organism	a character string specifying the scientific name of a query organism. Default is organism = NULL. In this case all available biomart connections are returned.
update	a logical value specifying whether or not the local listMart.txt and listDatasets.txt files shall be updated by remote access to BioMart.
mute_citation	logical value indicating whether citation message should be muted.

### Details

This function collects all available biomart connections and returns a table storing the organism for which biomart connections are available as well as the corresponding mart and database.

### Note

When you run this function for the first time, the data retrieval procedure will take some time, due to the remote access to BioMart. The corresponding result is then saved in a \*.txt file named "\_biomart/listDatasets.txt" in the `tempdir` directory, allowing subsequent queries to perform much faster.

### Author(s)

Hajk-Georg Drost

## References

<http://biomart.org/>

Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman, Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).

BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma and Wolfgang Huber, Bioinformatics 21, 3439-3440 (2005).

## See Also

[getMarts](#), [getDatasets](#), [biomart](#), [organismFilters](#), [organismAttributes](#)

## Examples

```
## Not run:
# returning all available biomart connections
head(organismBM(), 20)
# retrieving all available datasets and biomart connections for
# a specific query organism (scientific name)
organismBM(organism = "Homo sapiens")
# you can also update the downloaded version using
# the "update = TRUE" argument
head(organismBM(update = TRUE), 20)

## End(Not run)
```

---

organismFilters

*Retrieve Ensembl Biomart filters for a query organism*

---

## Description

In addition to the [organismBM](#) and [organismAttributes](#) functions, this function returns all available filters that can be accessed through different marts and datasets for a given query organism.

## Usage

```
organismFilters(organism, update = FALSE, topic = NULL)
```

## Arguments

organism	a character string specifying the scientific name of a query organism.
update	a logical value specifying whether or not the local listMart.txt, listDatasets.txt, and listFilters_organism.txt files shall be updated by remote access to BioMart.
topic	a character string specifying a topic (category) of filters, e.g. topic = "id".

## Details

For a given query organism, this function retrieves all available filters that can be accessed through different marts and datasets.

Sometimes the same filter names correspond to different datasets and marts causing problems when using `getMarts`. The approach introduced by this function provides (again) a organism centric way of accessing organism specific filters.

The `topic` argument allows the user to search for specific filters topics/categories for faster selection.

## Value

a data.frame storing corresponding filter names, description, datasets, and marts.

## Note

When you run this function for the first time, the data retrieval procedure will take some time, due to the remote access to BioMart. The corresponding result is then saved in a \*.txt file within the `tempdir` directory named "\_biomart/listMarts.txt", "\_biomart/listDatasets.txt", and "\_biomart/listFilters\_organism.txt", allowing subsequent queries to perform much faster.

## Author(s)

Hajk-Georg Drost

## References

<http://biomart.org/>

Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package `biomaRt`. Steffen Durinck, Paul T. Spellman, Ewan Birney and Wolfgang Huber, *Nature Protocols* 4, 1184-1191 (2009).

BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma and Wolfgang Huber, *Bioinformatics* 21, 3439-3440 (2005).

## See Also

[organismBM](#), [organismAttributes](#), [getAttributes](#), [getDatasets](#), [getMarts](#)

## Examples

```
## Not run:  
# search for filter topic "id"  
head(organismFilters("Homo sapiens", topic = "id"), 20)  
  
## End(Not run)
```

---

read\_assemblystats      *Import Genome Assembly Stats File*

---

### Description

This function reads an organism specific Genome Assembly Stats file that was retrieved with [getAssemblyStats](#).

### Usage

```
read_assemblystats(file, type = "raw")
```

### Arguments

file	a character string specifying the path to the file storing the Genome Assembly Stats file.
type	a tibble object, either type = "raw" to import the entire genome assembly stats file or type = "stats" to import overall statistics including all chromosomes, mitochondria and plastids.

### Details

This function takes a string specifying the path to the Genome Assembly Stats file of interest (e.g. the path returned by [getAssemblyStats](#)) and imports it.

### Author(s)

Hajk-Georg Drost

### See Also

[getAssemblyStats](#), [read\\_genome](#), [read\\_proteome](#), [read\\_cds](#), [read\\_gff](#)

---

read\_cds      *Import CDS as Biostrings or data.table object*

---

### Description

This function reads an organism specific CDS stored in a defined file format.

## Usage

```
read_cds(  
  file,  
  format = "fasta",  
  obj.type = "Biostrings",  
  delete_corrupt = FALSE,  
  ...  
)
```

## Arguments

file	a character string specifying the path to the file storing the CDS.
format	a character string specifying the file format used to store the genome, e.g. format = "fasta" (default) or format = "gbk".
obj.type	a character string specifying the object stype in which the genomic sequence shall be represented. Either as obj.type = "Biostrings" (default) or as obj.type = "data.table".
delete_corrupt	a logical value specifying whether potential CDS sequences that cannot be divided by 3 shall be excluded from the the dataset. Default is delete_corrupt = FALSE.
...	additional arguments that are used by <a href="#">read.fasta</a> .

## Details

The `read.cds` function takes a string specifying the path to the cds file of interest as first argument.

It is possible to read in different proteome file standards such as *fasta* or *genbank*.

CDS stored in fasta files can be downloaded from <http://www.ensembl.org/info/data/ftp/index.html>.

## Value

A `data.table` storing the gene id in the first column and the corresponding sequence as string in the second column.

## Author(s)

Hajk-Georg Drost

## See Also

Other cds: [getCDSSet\(\)](#), [getCDS\(\)](#)

Other readers: [read\\_genome\(\)](#), [read\\_gff\(\)](#), [read\\_proteome\(\)](#), [read\\_rna\(\)](#)



---

`read_genome`*Import Genome Assembly as Biostrings or data.table object*

---

## Description

This function reads an organism specific genome stored in a defined file format.

## Usage

```
read_genome(file, format = "fasta", obj.type = "Biostrings", ...)
```

## Arguments

<code>file</code>	a character string specifying the path to the file storing the genome.
<code>format</code>	a character string specifying the file format used to store the genome, e.g. <code>format = "fasta"</code> (default) or <code>format = "gbk"</code> .
<code>obj.type</code>	a character string specifying the object type in which the genomic sequence shall be represented. Either as <code>obj.type = "Biostrings"</code> (default) or as <code>obj.type = "data.table"</code> .
<code>...</code>	additional arguments that are used by the <a href="#">read.fasta</a> function.

## Details

This function takes a string specifying the path to the genome file of interest as first argument (e.g. the path returned by [getGenome](#)).

## Value

Either a `Biostrings` or `data.table` object.

## Author(s)

Hajk-Georg Drost

## See Also

Other genome: [getGenomeSet\(\)](#), [getGenome\(\)](#)

Other readers: [read\\_cds\(\)](#), [read\\_gff\(\)](#), [read\\_proteome\(\)](#), [read\\_rna\(\)](#)

---

`read_gff`*Import GFF File*

---

**Description**

This function reads an organism specific CDS stored in a defined file format.

**Usage**

```
read_gff(file)
```

**Arguments**

`file` a character string specifying the path to the file storing the CDS.

**Details**

This function takes a string specifying the path to the GFF file of interest (e.g. the path returned by [getGFF](#)).

**Value**

Either a `Biostrings` or `data.table` object.

**Author(s)**

Hajk-Georg Drost

**See Also**

Other `gff`: [getGFFSet\(\)](#), [getGFF\(\)](#)

Other readers: [read\\_cds\(\)](#), [read\\_genome\(\)](#), [read\\_proteome\(\)](#), [read\\_rna\(\)](#)

---

`read_proteome`*Import Proteome as Biostrings or data.table object*

---

**Description**

This function reads an organism specific proteome stored in a defined file format.

**Usage**

```
read_proteome(file, format = "fasta", obj.type = "Biostrings", ...)
```

**Arguments**

file	a character string specifying the path to the file storing the proteome.
format	a character string specifying the file format used to store the genome, e.g. format = "fasta" (default) or format = "gbk".
obj.type	a character string specifying the object stype in which the genomic sequence shall be represented. Either as obj.type = "Biostrings" (default) or as obj.type = "data.table".
...	additional arguments that are used by <a href="#">read.fasta</a> .

**Details**

This function takes a string specifying the path to the proteome file of interest as first argument. It is possible to read in different proteome file standards such as *fasta* or *genebank*.

**Value**

Either a Biostrings or data.table object.

**Author(s)**

Hajk-Georg Drost

**See Also**

Other readers: [read\\_cds\(\)](#), [read\\_genome\(\)](#), [read\\_gff\(\)](#), [read\\_rna\(\)](#)  
 Other proteome: [getProteomeSet\(\)](#), [getProteome\(\)](#)

---

read_rm	<i>Import Repeat Masker output file</i>
---------	---

---

**Description**

This function reads an organism specific Repeat Masker output file.

**Usage**

```
read_rm(file)
```

**Arguments**

file	a character string specifying the path to the file storing the Repeat Masker output (e.g. retrieved with <a href="#">getRepeatMasker</a> ).
------	---

**Details**

This function takes a string specifying the path to the Repeat Masker output file of interest as first argument.

**Author(s)**

Hajk-Georg Drost

**See Also**[getRepeatMasker](#), [read\\_genome](#), [read\\_proteome](#), [read\\_gff](#), [read\\_rna](#)

---

`read_rna`*Import RNA as Biostrings or data.table object*

---

**Description**

This function reads an organism specific RNA stored in a defined file format.

**Usage**

```
read_rna(file, format = "fasta", obj.type = "Biostrings", ...)
```

**Arguments**

<code>file</code>	a character string specifying the path to the file storing the RNA.
<code>format</code>	a character string specifying the file format used to store the genome, e.g. <code>format = "fasta"</code> (default) or <code>format = "gbk"</code> .
<code>obj.type</code>	a character string specifying the object type in which the genomic sequence shall be represented. Either as <code>obj.type = "Biostrings"</code> (default) or as <code>obj.type = "data.table"</code> .
<code>...</code>	additional arguments that are used by <a href="#">read.fasta</a> .

**Details**

This function takes a string specifying the path to the RNA file of interest as first argument. It is possible to read in different proteome file standards such as *fasta* or *genebank*.

**Value**

A `data.table` storing the gene id in the first column and the corresponding sequence as string in the second column.

**Author(s)**

Hajk-Georg Drost

**See Also**Other rna: [getRNASet\(\)](#), [getRNA\(\)](#)Other readers: [read\\_cds\(\)](#), [read\\_genome\(\)](#), [read\\_gff\(\)](#), [read\\_proteome\(\)](#)

---

refseqOrganisms	<i>Retrieve All Organism Names Stored on refseq</i>
-----------------	---

---

**Description**

This function extracts all organism names (scientific names) for which genomes, proteomes, and CDS files are stored on the NCBI refseq server.

**Usage**

```
refseqOrganisms()
```

**Author(s)**

Hajk-Georg Drost

---

summary_cds	<i>Retrieve summary statistics for a coding sequence (CDS) file</i>
-------------	---

---

**Description**

A summary statistics of specific CDS features is returned.

**Usage**

```
summary_cds(file, organism)
```

**Arguments**

file	file path to a CDS file in fasta format.
organism	character string specifying the organism at hand.

**Details**

The summary statistics include:

- total\_seqs:
- nnn\_abs: The total number of NNN's (over all chromosomes/scaffolds/contigs) in all coding sequences combined
- nnn\_perc: The percentage (relative frequency) of NNN's (over all chromosomes/scaffolds/contigs) compared to the total number of nucleotides of all coding sequences

**Author(s)**

Hajk-Georg Drost

**See Also**

[getCollection](#), [getCDS](#), [read\\_cds](#), [summary\\_genome](#)

---

summary\_genome

*Retrieve summary statistics for a genome assembly file*

---

**Description**

A summary statistics of specific genome features is generated. These statistics are useful to assess the genome quality of retrieved genome assemblies when performing comparative genomics tasks. This way, users can assess whether or not patterns found based on genome comparisons aren't just a technical artifact of differences in genome assembly quality.

**Usage**

```
summary_genome(file, organism)
```

**Arguments**

file	file path to a genome assembly file in fasta format.
organism	character string specifying the organism at hand.

**Details**

The summary statistics include:

- genome\_size\_mbp: Genome size in mega base pairs
- n50\_mbp: The N50 contig size of the genome assembly in mega base pairs
- n\_seqs: The number of chromosomes/scaffolds/contigs of the genome assembly file
- n\_nnn: The absolute number of NNNs (over all chromosomes or scaffolds or contigs) in the genome assembly file
- rel\_nnn: The percentage (relative frequency) of NNNs (over all chromosomes or scaffolds or contigs) compared to the total number of nucleotides in the genome assembly file
- genome\_entropy: The Shannon Entropy of the genome assembly file (median entropy over all individual chromosome entropies)
- n\_gc: The total number of GCs (over all chromosomes or scaffolds or contigs) in the genome assembly file
- rel\_gc: The (relative frequency) of GCs (over all chromosomes or scaffolds or contigs) compared to the total number of nucleotides in the genome assembly file

**Author(s)**

Hajk-Georg Drost

**See Also**

[summary\\_cds](#), [getCollection](#), [getGenome](#), [read\\_genome](#)

**Examples**

```
## Not run:  
# retrieve genome from NCBI RefSeq  
Sc <- biomart::getGenome(db = "refseq", organism = "Saccharomyces cerevisiae")  
# compute genome assembly summary statistics  
Sc_genome_summary <- summary_genome(file = Sc, organism = "Saccharomyces cerevisiae")  
# look at results  
Sc_genome_summary  
  
## End(Not run)
```

# Index

- \* **cachedir**
  - cachedir, [6](#)
  - cachedir\_set, [6](#)
- \* **cds**
  - getCDS, [18](#)
  - getCDSSet, [20](#)
  - read\_cds, [79](#)
- \* **collection**
  - getCollection, [22](#)
  - getCollectionSet, [24](#)
- \* **genome**
  - getGenome, [32](#)
  - getGenomeSet, [35](#)
  - read\_genome, [81](#)
- \* **getBioSet**
  - getBioSet, [16](#)
  - getCDSSet, [20](#)
  - getCollectionSet, [24](#)
  - getGenomeSet, [35](#)
  - getGFFSet, [39](#)
  - getProteomeSet, [52](#)
  - getRNASet, [59](#)
- \* **getBio**
  - getBio, [14](#)
  - getCDS, [18](#)
  - getCollection, [22](#)
  - getGenome, [32](#)
  - getGFF, [37](#)
  - getProteome, [50](#)
  - getRNA, [57](#)
- \* **gff**
  - getGFF, [37](#)
  - getGFFSet, [39](#)
  - read\_gff, [82](#)
- \* **meta\_retrieval**
  - meta.retrieval, [70](#)
  - meta.retrieval.all, [73](#)
- \* **package**
  - biomartr-package, [3](#)
- \* **proteome**
  - getProteome, [50](#)
  - getProteomeSet, [52](#)
  - read\_proteome, [82](#)
- \* **readers**
  - read\_cds, [79](#)
  - read\_genome, [81](#)
  - read\_gff, [82](#)
  - read\_proteome, [82](#)
  - read\_rna, [84](#)
- \* **rna**
  - getRNA, [57](#)
  - getRNASet, [59](#)
  - read\_rna, [84](#)
- biomart, [4](#), [41](#), [42](#), [75](#), [77](#)
- biomartr (biomartr-package), [3](#)
- biomartr-package, [3](#)
- cachedir, [6](#), [7](#), [62](#)
- cachedir\_set, [6](#), [6](#)
- check\_annotation\_biomart, [7](#)
- data.frame, [72](#)
- download.database, [8](#), [9](#), [64](#), [65](#)
- download.database.all, [8](#), [9](#), [65](#)
- ensembl\_divisions, [10](#), [10](#), [11](#), [31](#)
- get.ensembl.info, [10](#), [31](#), [45](#)
- getAssemblyStats, [11](#), [71](#), [73](#), [79](#)
- getAttributes, [13](#), [27](#), [31](#), [47](#), [78](#)
- getBio, [14](#), [18](#), [20](#), [24](#), [34](#), [38](#), [44](#), [51](#), [58](#)
- getBioSet, [16](#), [22](#), [26](#), [36](#), [41](#), [54](#), [60](#)
- getBM, [5](#), [42](#)
- getCDS, [13](#), [16](#), [18](#), [22](#), [24](#), [34](#), [38](#), [42](#), [44](#), [46](#), [51](#), [56](#), [58](#), [71](#), [73](#), [80](#), [86](#)
- getCDSSet, [18](#), [20](#), [20](#), [26](#), [36](#), [41](#), [54](#), [60](#), [80](#)
- getCollection, [13](#), [16](#), [20](#), [22](#), [26](#), [34](#), [38](#), [44](#), [51](#), [56](#), [58](#), [86](#), [87](#)



- getCollectionSet, [18](#), [22](#), [24](#), [24](#), [36](#), [41](#), [54](#), [60](#)
- getDatasets, [14](#), [26](#), [31](#), [42](#), [47](#), [77](#), [78](#)
- getENSEMBL, [27](#)
- getENSEMBL.gtf, [28](#)
- getENSEMBL.Seq, [29](#)
- getENSEMBLGENOMESInfo, [30](#)
- getENSEMBLInfo, [11](#), [30](#)
- getFilters, [14](#), [27](#), [31](#), [42](#), [47](#)
- getGenome, [13](#), [16](#), [20](#), [24](#), [32](#), [36](#), [38](#), [42](#), [44](#), [46](#), [51](#), [56](#), [58](#), [71](#), [73](#), [81](#), [87](#)
- getGENOMEREPORT, [34](#)
- getGenomeSet, [18](#), [22](#), [26](#), [34](#), [35](#), [41](#), [54](#), [60](#), [81](#)
- getGFF, [13](#), [16](#), [20](#), [24](#), [34](#), [37](#), [41](#), [48](#), [51](#), [56](#), [58](#), [71](#), [73](#), [82](#)
- getGFFSet, [18](#), [22](#), [26](#), [36](#), [38](#), [39](#), [44](#), [54](#), [60](#), [82](#)
- getGO, [41](#)
- getGroups, [42](#), [46](#), [71](#)
- getGTF, [43](#), [71](#), [73](#)
- getKingdomAssemblySummary, [11](#), [31](#), [45](#), [49](#), [61](#)
- getKingdoms, [42](#), [46](#), [61](#), [67](#), [71](#)
- getMarts, [4](#), [14](#), [27](#), [31](#), [42](#), [47](#), [75](#), [77](#), [78](#)
- getMetaGenomeAnnotations, [47](#), [49](#)
- getMetaGenomes, [47](#), [48](#), [48](#), [69](#)
- getMetaGenomeSummary, [45](#), [49](#), [61](#), [69](#)
- getProteome, [13](#), [16](#), [20](#), [24](#), [34](#), [38](#), [42](#), [44](#), [46](#), [50](#), [54](#), [56](#), [58](#), [71](#), [73](#), [83](#)
- getProteomeSet, [18](#), [22](#), [26](#), [36](#), [41](#), [51](#), [52](#), [60](#), [83](#)
- getReleases, [54](#)
- getRepeatMasker, [55](#), [71](#), [73](#), [83](#), [84](#)
- getRNA, [13](#), [16](#), [20](#), [24](#), [34](#), [38](#), [44](#), [51](#), [56](#), [57](#), [60](#), [71](#), [73](#), [84](#)
- getRNASet, [18](#), [22](#), [26](#), [36](#), [41](#), [54](#), [58](#), [59](#), [84](#)
- getSummaryFile, [45](#), [49](#), [61](#)
- getUniProtInfo, [62](#)
- getUniProtSTATS, [62](#)
  
- is.genome.available, [63](#), [67](#), [68](#)
  
- listAttributes, [5](#), [75](#)
- listDatabases, [8](#), [9](#), [64](#)
- listGenomes, [63](#), [65](#), [67](#), [68](#)
- listGroups, [67](#), [68](#)
- listKingdoms, [67](#), [68](#)
- listMetaGenomes, [48](#), [49](#), [69](#)
  
- listNCBIDatabases, [9](#)
- listNCBIDatabases (listDatabases), [64](#)
  
- meta.retrieval, [13](#), [42](#), [46](#), [56](#), [70](#), [74](#)
- meta.retrieval.all, [72](#), [73](#)
  
- organismAttributes, [14](#), [27](#), [31](#), [47](#), [74](#), [77](#), [78](#)
- organismBM, [5](#), [14](#), [27](#), [31](#), [42](#), [47](#), [74](#), [75](#), [76](#), [77](#), [78](#)
- organismFilters, [5](#), [14](#), [27](#), [31](#), [42](#), [47](#), [75](#), [77](#), [77](#)
  
- read.fasta, [80](#), [81](#), [83](#), [84](#)
- read\_assemblystats, [13](#), [79](#)
- read\_cds, [20](#), [22](#), [79](#), [79](#), [81–84](#), [86](#)
- read\_genome, [34](#), [36](#), [79](#), [80](#), [81](#), [82–84](#), [87](#)
- read\_gff, [38](#), [41](#), [44](#), [79–81](#), [82](#), [83](#), [84](#)
- read\_proteome, [51](#), [54](#), [79–82](#), [82](#), [84](#)
- read\_rm, [56](#), [83](#)
- read\_rna, [58](#), [60](#), [80–84](#), [84](#)
- refseqOrganisms, [85](#)
  
- summary\_cds, [85](#), [87](#)
- summary\_genome, [32](#), [86](#), [86](#)
  
- tempdir, [7](#), [15](#), [17](#), [23](#), [25](#), [38](#), [40](#), [44](#), [75](#), [76](#), [78](#)