Package 'igblastr'

July 16, 2025

Title User-friendly R Wrapper to IgBLAST

Description The igblastr package provides functions to conveniently install and use a local IgBLAST installation from within R. IgBLAST is described at <https://pubmed.ncbi.nlm.nih.gov/23671333/>. Online IgBLAST: <https://www.ncbi.nlm.nih.gov/igblast/>.

biocViews Immunology, Immunogenetics, ImmunoOncology, CellBiology

URL https://bioconductor.org/packages/igblastr

BugReports https://github.com/HyrienLab/igblastr/issues

Version 0.99.11

License Artistic-2.0

Encoding UTF-8

Depends R (>= 4.2.0), tibble, Biostrings

Imports methods, utils, stats, tools, R.utils, curl, httr, xml2, rvest, xtable, jsonlite, S4Vectors, IRanges, GenomeInfoDb

Suggests parallel, testthat, knitr, rmarkdown, BiocStyle, ggplot2, dplyr, scales, ggseqlogo

SystemRequirements Perl (for install_IMGT_germline_db() only)

VignetteBuilder knitr

Collate utils.R long_to_wide_airr.R file-utils.R db-utils.R IMGT-utils.R AIRR-utils.R precompiled-igblast-utils.R cache-utils.R get_igblast_root.R edit_imgt_file.R igblast_info.R auxiliary-data-utils.R install_igblast.R make_blastdbs.R create_region_db.R create_germline_db.R create_c_region_db.R list_germline_dbs.R list_c_region_dbs.R install_IMGT_germline_db.R install_AIRR_germline_db.R augment_germline_db.R igblastn-args-utils.R outfmt7-utils.R igblastn.R bcr_browser.R OAS-utils.R zzz.R

git_url https://git.bioconductor.org/packages/igblastr

git_branch devel

git_last_commit 8c29488

git_last_commit_date 2025-06-19

Repository Bioconductor 3.22

Date/Publication 2025-07-16

Author Hervé Pagès [aut, cre] (ORCID: <https://orcid.org/0009-0002-8272-4522>),
Ollivier Hyrien [aut, fnd] (ORCID:
<https://orcid.org/0000-0003-1909-2542>),
Kellie MacPhee [ctb] (ORCID: <https://orcid.org/0009-0008-0993-4009>),
Michael Duff [ctb] (ORCID: <https://orcid.org/0009-0008-4279-0756>),
Jason Taylor [ctb]

Maintainer Hervé Pagès <hpages.on.github@gmail.com>

Contents

`augment_germline_db` *Add novel gene alleles to a germline db*

Description

Three functions to add novel V, D, or J gene alleles to a germline db.

Note that these functions can also be used to combine germline databases from two different organisms. See "COMBINE GERMLINE DATABASES FROM TWO ORGANISMS" in the Examples section below for how to do this.

Usage

```
augment_germline_db_V(db_name, novel_alleles, destdir=".", overwrite=FALSE)
augment_germline_db_D(db_name, novel_alleles, destdir=".", overwrite=FALSE)
augment_germline_db_J(db_name, novel_alleles, destdir=".", overwrite=FALSE)
```

Arguments

`db_name`	A single string that is the name of the cached germline db that contains the set of gene alleles to augment. Use `list_germline_dbs()` to list the cached germline dbs. The exact function used (i.e. `augment_germline_db_V()`, `augment_germline_db_D()`, or `augment_germline_db_J()`) determines the set of alleles to augment (i.e. alleles from the V, D, or J region).

`novel_alleles`	A single string that is the path to a FASTA file (possibly gz-compressed) where the novel alleles are stored. Alternatively, the novel alleles can be supplied as a *named* DNAStringSet object.
`destdir`	A single string that is the path to the "destination directory", that is, the directory where the augmented V-, D-, or J-region db is to be created. This directory will be created if it doesn't exist already. Note that, by default, the augmented region db will be created in the current directory.
`overwrite`	If the "destination directory" already contains a V-, D-, or J-region db, should it be overwritten?

Value

An invisible NULL.

See Also

- The `igblastn` function to run the `igblastn` *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- `list_germline_dbs` to list the cached germline dbs.
- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.

Examples

```
if (!has_igblast()) install_igblast()

query <- system.file(package="igblastr", "extdata",
                     "heavy_sequences.fasta")

use_c_region_db("_IMGT.human.IGH+IGK+IGL.202412")

## ------------------------------------------------------------------------
## USE HUMAN GERMLINE DATABASE FROM AIRR
## ------------------------------------------------------------------------

use_germline_db("_AIRR.human.IGH+IGK+IGL.202501")

AIRR_df <- igblastn(query)

## ------------------------------------------------------------------------
## ADD NOVEL V ALLELES
## ------------------------------------------------------------------------

## 'fake_human_V_alleles.fasta' contains made-up novel V alleles:
## - 2 novel alleles for gene IGHV1-8: IGHV1-8*fake1, IGHV1-8*fake2
## - 1 novel allele for gene IGHV4-61: IGHV4-61*fake
my_novel_V_alleles <- system.file(package="igblastr", "extdata",
                                  "novel_germline_alleles",
                                  "fake_human_V_alleles.fasta")

## Take a quick look at these novel V alleles:
readDNAStringSet(my_novel_V_alleles)

## Create a new V germline database that combines the V alleles
## from _AIRR.human.IGH+IGK+IGL.202501 with our novel V alleles:
```

```
myVdb_path <- file.path(tempdir(), "myVdb")
augment_germline_db_V("_AIRR.human.IGH+IGK+IGL.202501",
                      my_novel_V_alleles,
                      destdir=myVdb_path)

## To use this new augmented V germline database with igblastn(),
## supply its path via the 'germline_db_V' argument:
AIRR_df2 <- igblastn(query, germline_db_V=myVdb_path)

## ---------------------------------------------------------------------
## A QUICK COMPARISON BETWEEN 'AIRR_df' AND 'AIRR_df2'
## ---------------------------------------------------------------------

## Index of rows where "v_call" has changed between 'AIRR_df'
## and 'AIRR_df2':
idx <- which(AIRR_df$v_call != AIRR_df2$v_call)
idx  # 2 rows

AIRR_df[idx, c("v_call", "v_cigar", "v_identity")]

AIRR_df2[idx, c("v_call", "v_cigar", "v_identity")]

## Besides these 2 rows, all the other rows are the same:
stopifnot(all.equal(AIRR_df[-idx, ], AIRR_df2[-idx, ]))

## ---------------------------------------------------------------------
## COMBINE GERMLINE DATABASES FROM TWO ORGANISMS
## ---------------------------------------------------------------------

## The augment_germline_db_[VDJ]() functions can be used to combine
## germline databases from two different organisms. This can be useful
## for example when working with BCR sequences from mice that have been
## engineered to have both mouse and some human immunoglobulin genes.
##
## To create a hybrid human/mouse V germline database, we can either:
##
## (1) Add all (or a subset of) mouse V alleles to all human V alleles.
##     This is done by extracting mouse V germline allele sequences from
##     a cached germline database and using them to augment a cached
##     germline database for human.
##
## (2) Add all (or a subset of) human V alleles to all mouse V alleles.
##     This is done by extracting human V germline allele sequences from
##     a cached germline database and using them to augment a cached
##     germline database for mouse.
##
## Note that:
## - We can choose to subset or not the V germline allele sequences
##   extracted from one V germline database before adding them to the
##   other V germline database.
## - The two approaches above are equivalent if we don't subset, that
##   is, if we combine **all** human V alleles with **all** mouse V
##   alleles.
## - However if our engineered mice only have a small known subset of
##   human immunoglobulin genes (e.g. IGHV1-2), then we might want to
##   create a hybrid human/mouse germline database that only adds the
##   human alleles for genes IGHV1-2 to the mouse V alleles. In this
```

```
##   case we need to use (2).

## Let's do (2):

db_name1 <- "_AIRR.mouse.PWD_PhJ.IGH+IGK+IGL.202501"
db_name2 <- "_AIRR.human.IGH+IGK+IGL.202501"

## Extract human V germline alleles:
human_V_alleles <- load_germline_db(db_name2, "V")

## Subset to keep only alleles for genes IGHV1-2:
idx <- grep("^IGHV[12]", names(human_V_alleles))
human_V12_alleles <- human_V_alleles[idx]

## Create a new V germline database that combines the mouse V
## alleles from 'db_name1' with the alleles in 'human_V12_alleles':
engmouseVdb_path <- file.path(tempdir(), "engmouseVdb")
augment_germline_db_V(db_name1, human_V12_alleles,
                      destdir=engmouseVdb_path)

## Then, assuming that 'query' contains BCR sequences from the
## engineered mice:
## Not run:
  use_germline_db(db_name1)
  use_c_region_db("_IMGT.mouse.IGH.202412")
  igblastn(query, germline_db_V=engmouseVdb_path, ...)

## End(Not run)

## Note that, by default, the mouse-only D and J databases that
## we selected with 'use_germline_db(db_name1)' are being used.
## If we also want to create hybrid D and J databases, we need
## to repeat the above steps for each of them. Then we need to
## specify the paths to the 3 hybrid databases when we call igblastn():
## Not run:
  igblastn(query, germline_db_V=engmouseVdb_path,
                  germline_db_D=engmouseDdb_path,
                  germline_db_J=engmouseJdb_path,
                  ...)

## End(Not run)
```

`auxiliary-data-utils` *Manipulation of IgBLAST auxiliary data*

Description

A standard IgBLAST installation – like the one used by the **igblastr** package – typically includes *auxiliary data* for various organisms, in the form of one tabulated file per organism. Each file indicates the germline J gene coding frame start position, the J gene type, and the CDR3 end position for each sequence in the germline J sequence database. See https://ncbi.github.io/igblast/cook/How-to-set-up.html for the details.

You can use `get_igblast_auxiliary_data()` to obtain the path to the file containing the auxiliary data for a given organism.

Usage

```
get_igblast_auxiliary_data(organism)
load_igblast_auxiliary_data(organism)
```

Arguments

organism	A single string containing the name of an organism as returned by list_igblast_organisms().

Value

get_igblast_auxiliary_data(): Returns a single string containing the path to the auxiliary data included in the IgBLAST installation used by **igblastr**, for the specified organism. Not necessarily suitable to use with igblastn() (see WARNING below).

load_igblast_auxiliary_data(): Returns the auxiliary data in a data.frame with 1 row per germline J sequence and the following columns:

1. sseqid: gene/allele name a.k.a. subject sequence id;
2. coding_frame_start: first coding frame start position (position is 0-based);
3. chaintype: chain type;
4. CDR3_stop: CDR3 stop;
5. extra_bps: extra base pairs beyond J coding end.

WARNING

According to https://ncbi.github.io/igblast/cook/How-to-set-up.html the auxiliary data included in IgBLAST is specific to a particular NCBI or IMGT germline db. Unfortunately this means that this data is NOT guaranteed to be compatible with the germline db that you will use with igblastn(). See documentation of the auxiliary_data argument in ?igblastn for more information about this.

See Also

- https://ncbi.github.io/igblast/cook/How-to-set-up.html for important information about the IgBLAST auxiliary data.
- The igblastn function to run the igblastn *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- install_igblast to perform an *internal* IgBLAST installation.
- get_igblast_root to get (or set) the IgBLAST installation used (or to be used) by the **igblastr** package.
- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.

Examples

```
if (!has_igblast()) install_igblast()

igblast_info()

list_igblast_organisms()

## Make sure to read the WARNING above before using the auxiliary
## data below with igblastn()!
```

```
get_igblast_auxiliary_data("human")
load_igblast_auxiliary_data("human")

get_igblast_auxiliary_data("rhesus_monkey")
load_igblast_auxiliary_data("rhesus_monkey")
```

`bcr_browser` *Display annotated BCR sequences in the browser*

Description

Use `bcr_browser()` to display the annotated BCR sequences returned by `igblastn()` in the browser. For each sequence, the V, D, and J segments are shown as well as the FWR1-4 and CDR1-3 regions. Additionally, the C segments are shown if the C-region information is available.

Usage

```
bcr_browser(AIRR_df, show.full.sequence=FALSE, dna.coloring=TRUE,
            Vcolor="#FFDDD2", Dcolor="#CFC", Jcolor="#CEF", Ccolor="#EEC",
            FWRcolor="#C9D", CDRcolor="#EE4")
```

Arguments

`AIRR_df`	The AIRR-formatted data.frame or tibble returned by `igblastn()`. Note that calling `bcr_browser()` on a data.frame with thousands of rows is quite resource-intensive (it can even crash your browser!), so in this case we recommend subsetting the data.frame before passing it to `bcr_browser()` to keep the number of rows under 2000.
`show.full.sequence`	By default, the part of the BCR sequences upstream of the V region is not shown. Set `show.full.sequence` to `TRUE` to show it.
`dna.coloring`	Whether the nucleotides in the BCR sequences (`sequence` column in `AIRR_df`) should be colored or not.
`Vcolor`, `Dcolor`, `Jcolor`, `Ccolor`	The background colors of the V, D, J, and C segments of the BCR sequences. Note that the C segments are shown only if `AIRR_df` contains C-region information.
`FWRcolor`, `CDRcolor`	The background colors of the Framework Regions (FWR1-4) and Complementarity-Determining Regions (CDR1-3), respectively.

Value

0 or the error code returned by the internal call to `browseURL()`, invisibly.

See Also

- The `igblastn` function to run the `igblastn` *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- IgBLAST is described at `https://pubmed.ncbi.nlm.nih.gov/23671333/`.
- tibble objects implemented in the **tibble** package.

Examples

```
if (!has_igblast()) install_igblast()

query <- system.file(package="igblastr", "extdata",
                     "heavy_sequences.fasta")
use_germline_db("_AIRR.human.IGH+IGK+IGL.202501")

## ------------------------------------------------------------------------
## With C regions
## ------------------------------------------------------------------------

use_c_region_db("_IMGT.human.IGH+IGK+IGL.202412")
AIRR_df <- igblastn(query)
bcr_browser(AIRR_df)

## By default, the part of the sequences upstream of the V region is
## not shown. Use 'show.full.sequence=TRUE' to show the full sequences:
bcr_browser(AIRR_df, show.full.sequence=TRUE)

## ------------------------------------------------------------------------
## No C regions
## ------------------------------------------------------------------------

use_c_region_db("")
AIRR_df2 <- igblastn(query)
bcr_browser(AIRR_df2)
```

---

get_igblast_root *Control IgBLAST installation to use*

Description

Get (or set) the IgBLAST installation used (or to be used) by the **igblastr** package.

Usage

```
get_igblast_root()
set_igblast_root(version_or_path)
```

Arguments

`version_or_path`
: A single string that is either a version number (e.g. `"1.22.0"`) or the path to an IgBLAST installation.

Details

`set_igblast_root` can be used to set or change the path to the IgBLAST installation to use. This can be an *internal* or *external* installation.

In the former case, `version_or_path` should be the version of an existing *internal* installation. The setting will be persistent.

In the latter case, it should be the full path (absolute or relative) to the *root directory* of a valid *external* installation. Note that the setting won't be persistent i.e. it won't be remembered across R

sessions. See ?IGBLAST_ROOT for how to set the *external* IgBLAST installation to use in **igblastr** in a persistent manner.

Value

get_igblast_root() returns a single string containing the path to the *root directory* of the IgBLAST installation used by **igblastr**.

set_igblast_root() returns a single string containing the path to the *root directory* of the newly selected IgBLAST installation. The string is returned invisibly.

See Also

- The igblastn function to run the igblastn *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- install_igblast to perform an *internal* IgBLAST installation.
- igblast_info to collect basic information about the IgBLAST installation used by the **igblastr** package.
- IGBLAST_ROOT to set the *external* IgBLAST installation to be used by the **igblastr** package in a persistent manner.
- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.

Examples

```
if (!has_igblast()) install_igblast()

get_igblast_root()
```

---

igblastn *BLAST for Ig and TCR sequences*

Description

The igblastn() function is a wrapper to the igblastn *standalone executable* included in IgBLAST. This is the main function in the **igblastr** package.

Usage

```
igblastn(query, outfmt="AIRR",
         germline_db_V="auto", germline_db_V_seqidlist=NULL,
         germline_db_D="auto", germline_db_D_seqidlist=NULL,
         germline_db_J="auto", germline_db_J_seqidlist=NULL,
         organism="auto", c_region_db="auto", auxiliary_data="auto",
         ...,
         out=NULL, parse.out=TRUE,
         show.in.browser=FALSE, show.command.only=FALSE)

igblastn_help(long.help=FALSE, show.in.browser=FALSE)
```

Arguments

query	A character vector containing the paths to the input files (FASTA), or a *named* DNAStringSet object. If a character vector, then query must be of length >= 1 and each vector element must be the path to a FASTA file (possibly gz-compressed). In the context of IgBLAST, the DNA sequences in the FASTA files are referred to as *the query sequences*, and the sequence names found in the description lines of the FASTA records are referred to as *the query sequence ids*. Note that the query sequences are typically (but not always) stored in a single file, in which case query will be a single string. If more than one FASTA file is specified via query, then igblastn() will concatenate all the files together and pass the resulting file to the igblastn *standalone executable*. If query is a DNAStringSet object, then it must have names on it. These will be considered the query sequence ids.
outfmt	One of "AIRR", 3, 4, 7, or 19. "AIRR" is the default and is an alias for 19. outfmt can also be a string describing a customized format 7 e.g. "7 qseqid sseqid pident nident length score". See ?list_outfmt7_specifiers for more information about customizing format 7.
germline_db_V	"auto" (the default), or the path to a V-region db. Note that, by default (i.e. when germline_db_V is omitted or set to "auto"), igblastn() uses the V-region db that belongs to the cached germline db currently selected. See ?use_germline_db for how to select the cached germline db to use with igblastn().
germline_db_D	Same as germline_db_V but for the D-region db.
germline_db_J	Same as germline_db_V but for the J-region db.
germline_db_V_seqidlist, germline_db_D_seqidlist, germline_db_J_seqidlist	Restrict search of germline database to list of gene alleles. A list of gene alleles can be specified either as a character vector of gene allele identifiers (e.g. IGHV3-23*01, IGHV3-23*04, etc...) or as the path to a file containing the identifiers (one identifier per line). In the latter case, a file object must be passed to the germline_db_V_seqidlist, germline_db_D_seqidlist, or germline_db_J_seqidlist argument. The file object will typically be constructed with something like file("path/to/some/file").
organism	"auto" (the default), or the organism associated with the query sequences. Supported organisms include human, mouse, rat, rabbit and rhesus_monkey. Use list_igblast_organisms() to obtain this list programmatically. Note that, by default (i.e. when organism is omitted or set to "auto"), igblastn() infers the organism from the name of the cached germline db currently selected. See ?use_germline_db for how to select the cached germline db to use with igblastn().
c_region_db	"auto" (the default), NULL, or the path to a C-region db. Note that, by default (i.e. when c_region_db is omitted or set to "auto"), igblastn() uses the cached C-region db currently selected. See ?use_c_region_db for how to select the cached C-region db to use with igblastn().
auxiliary_data	"auto" (the default), or the path to a file containing the coding frame start positions for the sequences in the J-region db, or NULL.

Note that, by default (i.e. when auxiliary_data is omitted or set to "auto"), igblastn() uses one of the auxiliary data files included in the IgBLAST installation used by **igblastr**. More precisely, igblastn() uses get_igblast_auxiliary_data() internally to obtain the path to the organism-specific auxiliary data file.

IMPORTANT NOTES:

- Supplying auxiliary data that is not compatible with the V gene sequences of the selected germline db can cause igblastn() to return improper frame status or CDR3 information (other returned information will still be correct). See ?get_igblast_auxiliary_data for more information.
- When auxiliary_data is set to NULL, then no auxiliary data is used. In this case, igblastn() can emit a significant number of the following warning:

```
Warning: Auxilary data file could not be found
```

and various columns of the returned AIRR-formatted tibble (e.g. columns vj_in_frame, productive, cdr3, fwr4, and others) will be filled with NAs.

`...` Extra arguments to be passed to the igblastn *standalone executable*. The list of valid arguments can be displayed with igblastn_help().

Note that the argument/value pairs must be passed to the igblastn() function in the usual R fashion. For example, what would be passed as -num_alignments 1 -num_threads 8 when invoking the igblastn *standalone executable* in a terminal should be passed as num_alignments_V=1, num_threads=8 when calling the igblastn() function:

```
igblastn(query, num_alignments_V=1, num_threads=8)
```

`out` NULL (the default), or the path to the file where the igblastn *standalone executable* should write its output.

Note that, by default (i.e. when out is omitted or set to NULL), igblastn() instructs the igblastn *standalone executable* to write its output to a temporary file.

`parse.out` Whether igblastn() should parse the plain-text output produced by the igblastn *standalone executable* or not, before returning it to the user. TRUE by default.

If set to FALSE, then igblastn() returns the output as-is in a character vector, with one line per element in the vector. Note that igblastn() sets the "igblastn_raw_output" class attribute on this character vector, which allows compact display of the vector (this is achieved via a dedicated print() method defined in the **igblastr** package). The class attribute can be dropped with unclass().

`show.in.browser`

For igblastn(): Whether the output of the igblastn *standalone executable* should also be displayed in the browser or not (in addition to being returned by the igblastn() function call). FALSE by default.

For igblastn_help(): Whether the help printed by the igblastn *standalone executable* (when invoked with the -h or -help argument) should be displayed in the browser or not. FALSE by default.

`show.command.only`

TRUE or FALSE. If set to TRUE, igblastn() won't invoke the igblastn *standalone executable* and instead will display the full command that shows how it would have invoked it. Note that the command is also returned in an invisible character vector. FALSE by default.

`long.help` TRUE or FALSE. If set to FALSE (the default), the igblastn *standalone executable* is invoked with the -h argument. Otherwise, it's invoked with the -help argument.

Value

`igblastn()` captures the output produced by the `igblastn` *standalone executable* and returns it as:

- A tibble with 1 row per query sequence if `outfmt` is `"AIRR"` or `19` and `parse.out` is `TRUE`.
- A nested list with two top-level components (`records` and `footer`) if `outfmt` is `7` (or a customized format 7) and `parse.out` is `TRUE`. See `?parse_outfmt7` for more information.
- A character vector with class attribute `"igblastn_raw_output"` on it in all other cases, that is, if `parse.out` is `FALSE` or `outfmt` is `3` or `4`. See the `parse.out` argument above for more information.

Note

By default, the NCBI BLAST+ and IgBLAST programs will "call home" to report usage when they run on a computer with internet access. See https://www.ncbi.nlm.nih.gov/books/NBK569851/ for the details. This can induce a significant slowdown in some situations e.g. when the `igblastn` *standalone executable* is called in a loop on a small set of query sequences at each iteration.

For this reason, the "call home" feature is disabled in **igblastr** by default, unless environment variable BLAST_USAGE_REPORT is set to true. See `?igblastr_usage_report` for more information.

See Also

- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.
- `install_igblast` to perform an *internal* IgBLAST installation.
- `igblast_info` to collect basic information about the IgBLAST installation used by the **igblastr** package.
- `install_IMGT_germline_db` to install a germline db from IMGT.
- `use_germline_db` to select the cached germline db to use with `igblastn()`.
- `use_c_region_db` to select the cached C-region db to use with `igblastn()`.
- `bcr_browser` to display the annotated sequences returned by `igblastn()` in the browser.
- `list_outfmt7_specifiers` for how to customize output format 7.
- `list_igblast_organisms` to list the organisms supported by IgBLAST.
- `augment_germline_db` to add novel gene alleles to a germline db.
- igblastr_usage_report to turn "Usage Reporting" on or off.
- DNAStringSet objects implemented in the **Biostrings** package.
- tibble objects implemented in the **tibble** package.

Examples

```
if (!has_igblast()) install_igblast()

## ------------------------------------------------------------------------
## Preliminary steps
## ------------------------------------------------------------------------

igblast_info()

## Files 'heavy_sequences.fasta' and 'light_sequences.fasta' included
## in igblastr contain 250 paired heavy- and light- chain sequences (125
```

```
## sequences in each file) downloaded from OAS (the Observed Antibody
## Space database):
filenames <- paste0(c("heavy", "light"), "_sequences.fasta")
query <- system.file(package="igblastr", "extdata", filenames)

## Install Human germline db from IMGT (Perl required!):
db_name <- install_IMGT_germline_db("202506-1", "Homo_sapiens", force=TRUE)

## Select germline db to use with igblastn():
use_germline_db(db_name)

## Select C-region db to use with igblastn():
use_c_region_db("_IMGT.human.IGH+IGK+IGL.202412")

## ------------------------------------------------------------------------
## Call igblastn()
## ------------------------------------------------------------------------

## We don't specify the 'outfmt' argument so output will be in AIRR format.
AIRR_df <- igblastn(query)
AIRR_df

## The result is a tibble with one row per query sequence:
class(AIRR_df)
dim(AIRR_df)

## You can call bcr_browser() on 'AIRR_df' to vizualize the annotated
## sequences in the browser. See '?bcr_browser'.

## Note that this tibble can easily be converted to an ordinary data.frame
## with 'as.data.frame()', or to a DataFrame with 'as(., "DataFrame")':
as(AIRR_df, "DataFrame")

## ------------------------------------------------------------------------
## More examples
## ------------------------------------------------------------------------

## See '?parse_outfmt7' for more examples.
```

`igblastr_usage_report` *Turn "Usage Reporting" on or off*

Description

By default, the NCBI BLAST+ and IgBLAST programs will "call home" to report usage when they run on a computer with internet access. See https://www.ncbi.nlm.nih.gov/books/NBK569851/ for the details. This can induce a significant slowdown in some situations e.g. when the `igblastn` *standalone executable* is called in a loop on a small set of query sequences at each iteration.

For this reason, the "call home" feature is disabled in **igblastr** by default, unless environment variable BLAST_USAGE_REPORT is set to true.

More precisely, the "call home" feature is controlled by global option `igblastr_usage_report` in **igblastr**. On package startup, this option is set to `TRUE` if environment variable BLAST_USAGE_REPORT is set to true. Otherwise (i.e. if BLAST_USAGE_REPORT is not set, or is set to false or gibberish) it is set to `FALSE`.

Details

The user can change the value of global option `igblastr_usage_report` any time with:

```
options(igblastr_usage_report=TRUE)
```

or with:

```
options(igblastr_usage_report=FALSE)
```

To get the value of this option, use:

```
getOption("igblastr_usage_report")
```

Note that changing the value of a global option interactively with `options(...)` won't be remembered across R sessions. For a persistent change, you can either:

- Put the `options(...)` command in your .Rprofile file. See ?`Rprofile` for more information. Note that this is the standard way of setting global options persistently.
- In the particular case of global option `igblastr_usage_report` an alternative is to define environment variable BLAST_USAGE_REPORT outside R. The exact way to do this is OS-dependent e.g. on Linux and Mac you can define it in your user's `.profile` by adding the following line to it:

```
export BLAST_USAGE_REPORT=true
```

See Also

- The `igblastn` function to run the `igblastn` *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.

Examples

```
## Check current status of usage reporting:
getOption("igblastr_usage_report")

## Turn on usage reporting:
options(igblastr_usage_report=TRUE)

## Turn off usage reporting:
options(igblastr_usage_report=FALSE)
```

---

igblast_info *Check IgBLAST used by igblastr*

Description

Collect basic information about the IgBLAST installation used by the **igblastr** package, or about any IgBLAST installation on the user machine.

Usage

```
igblast_info(igblast_root=get_igblast_root())

igblast_build(igblast_root=get_igblast_root())
igblastn_version(igblast_root=get_igblast_root(), raw.version=FALSE)
makeblastdb_version(igblast_root=get_igblast_root(), raw.version=FALSE)
list_igblast_organisms(igblast_root=get_igblast_root())

has_igblast()
```

Arguments

igblast_root	A single string that is the path to an IgBLAST installation. By default igblast_root is set to get_igblast_root(), which is the path to the IgBLAST installation used by the **igblastr** package. See ?get_igblast_root for more information. Note that the supplied string must contain the path to the *root directory* of an IgBLAST installation, that is, to a directory with a bin subdirectory in it that has the igblastn, igblastp, and makeblastdb *standalone executables* (on Windows these executables are files named igblastn.exe, igblastp.exe, and makeblastdb.exe, respectively).

raw.version	By default (i.e. when raw.version is omitted or set to FALSE), igblastn_version() and makeblastdb_version() return the version string of the igblastn and makeblastdb *standalone executables* included in IgBLAST. This string is extracted from the output produced by system commands:
	`igblastn -version`
	and
	`makeblastdb -version`
	When raw.version is set to TRUE, igblastn_version() and makeblastdb_version() return the *full ouput* produced by the above commands.

Value

igblast_info() returns a named list containing basic information about the IgBLAST installation.

igblast_build() returns a single string containing IgBLAST build information.

By default, igblastn_version() returns a single string containing the version of the igblastn *standalone executable* included in IgBLAST.

By default, makeblastdb_version() returns a single string containing the version of the makeblastdb *standalone executable* included in IgBLAST.

list_igblast_organisms() returns a character vector that lists the organisms for which IgBLAST provides internal data. Note that this is obtained by simply listing the content of the internal_data directory located in the IgBLAST installation.

has_igblast() returns TRUE or FALSE.

See Also

- The igblastn function to run the igblastn *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- install_igblast to perform an *internal* IgBLAST installation.

- get_igblast_root to get (or set) the IgBLAST installation used (or to be used) by the **igblastr** package.
- IGBLAST_ROOT to set the *external* IgBLAST installation to be used by the **igblastr** package in a persistent manner.
- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.

Examples

```
if (!has_igblast()) install_igblast()

igblast_info()

list_igblast_organisms()
```

IGBLAST_ROOT *Use an external IgBLAST installation*

Description

Select the *external* IgBLAST installation to use in **igblastr** in a persistent manner.

Details

The **igblastr** package can use 2 types of IgBLAST installation:

1. Internal (a.k.a. igblastr-managed): refers to an installation obtained with `install_igblast()`.
2. External: refers to an installation that is not managed by the **igblastr** package. This is usually an installation that was manually performed by you or a system administrator on your machine. It can be a system-wide installation or a per-user installation.

To use an *external* installation of IgBLAST in **igblastr**, set environment variable IGBLAST_ROOT to the path of the installation. Note that this must be the path to the *root directory* of the IgBLAST installation, that is, to a directory with a `bin` subdirectory in it that has the `igblastn`, `igblastp`, and `makeblastdb` *standalone executables* (on Windows these executables are files named `igblastn.exe`, `igblastp.exe`, and `makeblastdb.exe`, respectively).

This can be done within your current R session with `Sys.setenv(IGBLAST_ROOT="path/to/igblast_root")` for testing. However, this won't be remembered across R sessions.

To set IGBLAST_ROOT in a persistent manner, define it outside R. The exact way to do this is OS-dependent e.g. on Linux and Mac you can define it in your user's `.profile` by adding the following line to it:

```
export IGBLAST_ROOT="path/to/igblast_root"
```

See Also

- The `igblastn` function to run the `igblastn` *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- `install_igblast` to perform an *internal* IgBLAST installation.
- `igblast_info` to collect basic information about the IgBLAST installation used by the **igblastr** package.
- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.

`install_igblast` *Install IgBLAST*

Description

Download and install a pre-compiled IgBLAST from NCBI FTP site for use with **igblastr**.

Usage

```
install_igblast(release="LATEST", force=FALSE, ...)
```

Arguments

`release`	A single string specifying the IgBLAST release version to install. For example `"LATEST"` (recommended), or one of the IgBLAST release versions listed at https://ftp.ncbi.nih.gov/blast/executables/igblast/release/ (e.g. `"1.21.0"`). Note that old versions have not been tested and are not guaranteed to be compatible with the **igblastr** package.
`force`	Set to `TRUE` to reinstall if the specified IgBLAST release version is already installed.
`...`	Extra arguments to be passed to the internal call to `download.file()`. See `?download.file` in the **utils** package for more information.

Value

The path to the *root directory* of the IgBLAST installation, as an invisible string.

See Also

- The `igblastn` function to run the `igblastn` *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- IGBLAST_ROOT to use an *external* IgBLAST installation.
- `igblast_info` to collect basic information about the IgBLAST installation used by the **igblastr** package.
- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.

Examples

```
if (!has_igblast()) install_igblast()

igblast_info()
```

install_IMGT_germline_db *Install a germline db from IMGT*

Description

The `install_IMGT_germline_db()` function downloads IMGT/V-QUEST germline sequences from the IMGT website for a given organism, and stores them in a local germline database. This local database gets installed in **igblastr**'s persistent cache. It can then be used later with `igblastn()`.

Usage

```
install_IMGT_germline_db(release, organism="Homo sapiens",
                         force=FALSE, ...)

## Related utilities:
list_IMGT_releases(recache=FALSE)
list_IMGT_organisms(release)
IMGT_is_up()
```

Arguments

`release`	A single string specifying the IMGT/V-QUEST release to get the germline sequences from (or to list the organisms from for `list_IMGT_organisms()`). Use `list_IMGT_releases()` to list all releases.
`organism`	A single string specifying the organism for which to get the germline sequences.
`force`	Set to `TRUE` to reinstall if the requested database is already installed.
`...`	Extra arguments to be passed to the internal call to `download.file()`. See `?download.file` in the **utils** package for more information.
`recache`	`list_IMGT_releases()` uses a caching mechanism so that the list of IMGT/V-QUEST releases gets downloaded only once from the IMGT website during an R session (note that this caching is done in memory so it does not persist across sessions). Set `recache` to `TRUE` to force a new download (and recaching) of the list of IMGT/V-QUEST releases.

Value

`install_IMGT_germline_db()` returns the name to the newly installed germline db as an invisible string.

`list_IMGT_releases()` returns the list of IMGT/V-QUEST releases in a character vector. The releases are sorted from newest to oldest (latest release is first).

`list_IMGT_organisms()` returns the list of organisms included in the specified IMGT/V-QUEST release in a character vector.

`IMGT_is_up()` returns `TRUE` or `FALSE`, indicating whether the IMGT website at `https://www.imgt.org` is up and running or down.

Note

install_IMGT_germline_db() generates the local database by performing the instructions provided at https://ncbi.github.io/igblast/cook/How-to-set-up.html.

In particular, utility tool *edit_imgt_file.pl* (Perl script included in the IgBLAST installation) is invoked to process the sequences. This means that Perl must be available on your machine.

See Also

- The igblastn function to run the igblastn *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- use_germline_db to select the cached germline db to use with igblastn().
- The IMGT website: https://www.imgt.org/.
- The IMGT/V-QUEST download site: https://www.imgt.org/download/V-QUEST/.
- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.

Examples

```
if (!has_igblast()) install_igblast()

if (IMGT_is_up()) {
  ## As of March 26, 2025, the latest IMGT/V-QUEST release is 202506-1:
  list_IMGT_releases()

  list_IMGT_organisms("202506-1")

  ## Download Mouse germline sequences from IMGT/V-QUEST 202506-1, and
  ## store them in a cached germline database:
  install_IMGT_germline_db("202506-1", organism="Mus musculus", force=TRUE)

  ## List the cached germline databases:
  list_germline_dbs()

  ## Tell igblastn() to use the newly installed germline db:
  use_germline_db("IMGT-202506-1.Mus_musculus.IGH+IGK+IGL")
}
```

---

list_c_region_dbs *List cached C-region dbs and select one to use with igblastn()*

Description

Use list_c_region_dbs() to list all the *cached C-region dbs*, that is, all the C-region databases currently installed in **igblastr**'s persistent cache.

Use use_c_region_db() to select the cached C-region db to use with igblastn(). This choice will be remembered for the duration of the current R session but can be changed anytime.

Use load_c_region_db() to load the nucleotide sequences of the gene regions stored in a cached C-region db.

Usage

```
list_c_region_dbs(builtin.only=FALSE, names.only=FALSE, long.listing=FALSE)

use_c_region_db(db_name=NULL, verbose=FALSE)

load_c_region_db(db_name)
```

Arguments

builtin.only	By default list_c_region_dbs() returns the list of all cached C-region dbs, including built-in C-region dbs. Set builtin.only to TRUE to return only the list of built-in C-region dbs. Note that built-in dbs are prefixed with an underscore (_).
names.only	By default list_c_region_dbs() returns the list of cached C-region dbs in a data.frame with one db per row. Set names.only to TRUE to return only the db names in a character vector.
long.listing	TRUE or FALSE. If set to TRUE, then list_c_region_dbs() returns a named list with one list element per C-region db. Each list element is a named integer vector that indicates the number of C-region sequences per locus. Ignored if names.only is set to TRUE.
db_name	For use_c_region_db(): NULL or a single string specifying the name of the cached C-region db to use. Use list_c_region_dbs() to list all the cached C-region dbs. If set to NULL (the default), then use_c_region_db() returns the name of the cached C-region db that is currently in use, if any. Otherwise it returns the empty string (""). Note that the current selection can be cancelled with use_c_region_db(""). For load_c_region_db(): A single string specifying the name of the cached C-region db from which to load the gene regions. Use list_c_region_dbs() to list all the cached C-region dbs.
verbose	If set to TRUE, then use_c_region_db() will display some information about its internal operations.

Details

The **igblastr** package provides a small set of utilities to manage the cached germline and C-region databases to use with igblastn().

Terminology:

- A *cached germline db* contains the nucleotide sequences of the V, D, and J gene regions for a given organism.
- A *cached C-region db* contains the nucleotide sequences of the C regions (i.e. constant gene regions) for a given organism.

This man page documents the basic utilities to operate on cached C-region dbs: list_c_region_dbs(), use_c_region_db(), and load_c_region_db().

The basic utilities to operate on cached germline dbs are documented in the man page for list_germline_dbs.

Value

`list_c_region_dbs()` returns the list of all cached C-region dbs in a data.frame with one db per row (if `names.only` is `FALSE`, which is the default), or in a character vector (if `names.only` is `TRUE`). Column C in the data.frame indicates the number of C-region sequences in each db.

Built-in dbs are prefixed with an underscore (_). Note that the C-region built-in dbs from IMGT were downloaded from https://www.imgt.org/vquest/refseqh.html#constant-sets and included in the **igblastr** package on the date indicated by the suffix of the db name.

When called with no argument, `use_c_region_db()` returns a single string containing the name of the cached C-region db currently used by `igblastn()` if any, or the empty string (`""`) if `igblastn()` is not using any C-region db.

When called with the `db_name` argument, `use_c_region_db(db_name)` returns `db_name` invisibly.

`load_c_region_db()` returns the nucleotide sequences from the specified C-region db in a named DNAStringSet object.

See Also

- The `igblastn` function to run the `igblastn` *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- `use_germline_db` to select the cached germline db to use with `igblastn()`.
- DNAStringSet objects in the **Biostrings** package.
- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.

Examples

```
if (!has_igblast()) install_igblast()

## 3 built-in C-region dbs prefixed with an underscore:
list_c_region_dbs()

list_c_region_dbs(names.only=TRUE)  # db names only

list_c_region_dbs(long.listing=TRUE)  # long listing

## Select C-region db to use with igblastn():
use_c_region_db("_IMGT.human.IGH+IGK+IGL.202412")

use_c_region_db()  # get current selection

use_c_region_db("")  # cancel current selection

use_c_region_db()

## Load C-region sequences:
load_c_region_db("_IMGT.human.IGH+IGK+IGL.202412")
load_c_region_db("_IMGT.mouse.IGH.202412")
```

`list_germline_dbs` *List cached germline dbs and select one to use with igblastn()*

Description

Use `list_germline_dbs()` to list all the *cached germline dbs*, that is, all the germline databases currently installed in **igblastr**'s persistent cache.

Use `use_germline_db()` to select the cached germline db to use with `igblastn()`. This choice will be remembered for the duration of the current R session but can be changed anytime.

Use `load_germline_db()` to load the nucleotide sequences of the gene regions stored in a cached germline db.

Usage

```
list_germline_dbs(builtin.only=FALSE, names.only=FALSE, long.listing=FALSE)

use_germline_db(db_name=NULL, verbose=FALSE)

load_germline_db(db_name, region_types=NULL)
```

Arguments

`builtin.only`	By default `list_germline_dbs()` returns the list of all cached germline dbs, including built-in germline dbs. Set `builtin.only` to `TRUE` to return only the list of built-in germline dbs. Note that built-in dbs are prefixed with an underscore (_).
`names.only`	By default `list_germline_dbs()` returns the list of cached germline dbs in a data.frame with one db per row. Set `names.only` to `TRUE` to return only the db names in a character vector.
`long.listing`	`TRUE` or `FALSE`. If set to `TRUE`, then `list_germline_dbs()` returns a named list with one list element per germline db. Each list element is an integer matrix that indicates the number of germline sequences per locus and region type. Ignored if `names.only` is set to `TRUE`.
`db_name`	For `use_germline_db()`: `NULL` or a single string specifying the name of the cached germline db to use. Use `list_germline_dbs()` to list all the cached germline dbs. If set to `NULL` (the default), then `use_germline_db()` returns the name of the cached germline db that is currently in use, if any. Otherwise it raises an error. For `load_germline_db()`: A single string specifying the name of the cached germline db from which to load the V, D, and/or J regions. Use `list_germline_dbs()` to list all the cached germline dbs.
`verbose`	If set to `TRUE`, then `use_germline_db()` will display some information about its internal operations.
`region_types`	The types of regions (V, D, and/or J) to load from the database. Specified as a single string (e.g. `"DJ"`) or as a character vector of single-letter elements (e.g. `c("D", "J")`). By default (i.e. when `region_types` is `NULL`), all the regions are returned.

Details

The **igblastr** package provides a small set of utilities to manage the cached germline and C-region databases to use with `igblastn()`.

Terminology:

- A *cached germline db* contains the nucleotide sequences of the V, D, and J gene regions for a given organism.
- A *cached C-region db* contains the nucleotide sequences of the C regions (i.e. constant gene regions) for a given organism.

This man page documents the basic utilities to operate on cached germline dbs: `list_germline_dbs()`, `use_germline_db()`, and `load_germline_db()`.

The basic utilities to operate on cached C-region dbs are documented in the man page for `list_c_region_dbs`.

Value

`list_germline_dbs()` returns the list of all cached germline dbs in a data.frame with one db per row (if `names.only` is `FALSE`, which is the default), or in a character vector (if `names.only` is `TRUE`). Columns V, D, J in the data.frame indicate the number of germline sequences for each region in each db.

Built-in dbs are prefixed with an underscore (_). Note that the germline built-in dbs from AIRR were obtained from https://ogrdb.airr-community.org/germline_sets/Homo%20sapiens and https://ogrdb.airr-community.org/germline_sets/Mus%20musculus and included in the **igblastr** package on the date indicated by the suffix of the db name.

When called with no argument, `use_germline_db()` returns a single string containing the name of the cached germline db currently used by `igblastn()` if any, or it raises an error if no germline db has been selected yet.

When called with the `db_name` argument, `use_germline_db(db_name)` returns `db_name` invisibly.

`load_germline_db()` returns the nucleotide sequences from the specified germline db in a named DNAStringSet object.

See Also

- The `igblastn` function to run the `igblastn` *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- `install_IMGT_germline_db` to install a germline db from IMGT.
- `use_c_region_db` to select the cached C-region db to use with `igblastn()`.
- DNAStringSet objects in the **Biostrings** package.
- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.

Examples

```
if (!has_igblast()) install_igblast()

## Get list of built-in germline dbs only.
list_germline_dbs(builtin.only=TRUE)
list_germline_dbs(builtin.only=TRUE, names.only=TRUE)  # db names only

## Long listing:
list_germline_dbs(long.listing=TRUE)
```

```
if (IMGT_is_up()) {
  ## Install Mouse germline db from IMGT (Perl required!):
  install_IMGT_germline_db("202506-1", "Mus_musculus", force=TRUE)

  list_germline_dbs()  # all germline dbs

  ## Select germline db to use with igblastn():
  db_name <- "IMGT-202506-1.Mus_musculus.IGH+IGK+IGL"
  use_germline_db(db_name)  # select germline db to use

  use_germline_db()  # get current selection

  ## Load germline sequences:
  load_germline_db(db_name)
  load_germline_db(db_name, region_types="D")
  load_germline_db(db_name, region_types="DJ")
}
```

`OAS-utils` *Download and manipulate OAS data*

Description

Some utility functions to query the Observed Antibody Space database, a.k.a. OAS, and to download and manipulate data from OAS.

OAS's homepage: https://opig.stats.ox.ac.uk/webapps/oas/

Note that OAS has two databases: the "Unpaired Sequences" database and the "Paired Sequences" database. Some of the utilities documented in this man page only work on data coming from the latter.

Usage

```
## Read metadata/data from a single OAS unit file:
read_OAS_csv_metadata(file)
read_OAS_csv(file, skip=1, ...)
extract_sequences_from_paired_OAS_df(df, add.prefix=FALSE)

## Basic query of OAS website:
list_paired_OAS_studies(as.df=FALSE, recache=FALSE)
list_paired_OAS_units(study, as.df=FALSE, recache=FALSE)
download_paired_OAS_units(study, units=NULL, destdir=".", ...)

## Read metadata/data from a batch of downloaded OAS unit files:
extract_metadata_from_OAS_units(dir=".", pattern="\\.csv\\.gz$")
extract_sequences_from_paired_OAS_units(dir=".", pattern="\\.csv\\.gz$")
```

Arguments

`file`	A single string that is the path to an *OAS unit file*.
`skip`	The number of lines of the data file to skip before beginning to read data. The first line in an OAS unit file contains metadata in JSON format, so must always be skipped.

...	For read_OAS_csv(): Extra arguments to be passed to the internal call to read.table(). See ?read.table in the **utils** package for more information. For download_paired_OAS_units(): Extra arguments to be passed to the internal call to download.file(). See ?download.file in the **utils** package for more information.
df	The data.frame or tibble returned by read_OAS_csv().
add.prefix	TRUE or FALSE. Should the names on the returned DNAStringSet object be the original sequence ids as-is (this is the default), or should the heavy_chain_ and light_chain_ prefixes be added to them? extract_sequences_from_paired_OAS_df() returns a DNAStringSet object with the sequence ids as names. The sequence ids are obtained from the sequence_id_heavy and sequence_id_light columns of the supplied data.frame or tibble. By default, they are propagated as-is to the DNAStringSet object, which makes it difficult to recognize which chain (heavy or light) the antibody sequences are coming from. Setting add.prefix to TRUE will add the heavy_chain_ or light_chain_ prefix to the names on the DNAStringSet object, hence making it easy to identify which chain a given antibody sequence is coming from.
as.df	TRUE or FALSE. By default, i.e. when as.df is FALSE, list_paired_OAS_studies() and list_paired_OAS_units() return the list of studies or units in a character vector. Alternatively you can set as.df to TRUE to get the list in a 3-column data.frame that contains a directory index as displayed at https://opig.stats.ox.ac.uk/webapps/ngsdb/paired/ or at https://opig.stats.ox.ac.uk/webapps/ngsdb/paired/Jaffe_2022/csv/.
recache	TRUE or FALSE. list_paired_OAS_studies() and list_paired_OAS_units() both cache the information retrieved from OAS website for the duration of the R session (note that this caching is done in memory so it does not persist across sessions). Set recache to TRUE to force a new retrieval (and recaching) of the results.
study	A single string containing the name of a study as returned by list_paired_OAS_studies().
units	NULL, or a character vector that must be a subset of list_paired_OAS_units(study) in which case the download will be restricted to these units only.
destdir	A single string that is the path to the directory where the OAS unit files are to be downloaded.
dir	A single string that is the path to a directory containing OAS unit files. This will typically be the same as destdir above if the unit files were downloaded with download_paired_OAS_units().
pattern	Regular expression passed to the internal call to list.files() to obtain the list of OAS unit files located in dir. No reason to change this unless you know what you are doing.

Details

OAS delivers data in the form of *OAS unit files*. These files are typically obtained by running the bulk_download.sh script that OAS generates based on one's search criteria. They are compressed CSV (comma-separated values) files with the .csv.gz extension.

OAS unit files can vary a lot in size: from only a few KB to 25 MB or more.

The first line in an OAS unit file contains metadata in JSON format (which means that these files cannot strictly be considered CSV files).

The CSV data is MiAIRR-compliant (see The "MiAIRR format" paper in the References section below).

Value

`read_OAS_csv_metadata()` extracts the metadata from the specified OAS unit file and returns it in a named list.

`read_OAS_csv()` extracts the data from the specified OAS unit file and returns it in a tibble. The tibble has 1 row per antibody sequence if the data is unpaired (i.e. comes from the "Unpaired Sequences" database), or 1 row per sequence pair if the data is paired (i.e. comes from the "Paired Sequences" database).

`extract_sequences_from_paired_OAS_df()` returns the sequence pairs in a named DNAStringSet object where the names are the sequence ids. See `add.prefix` above for how the sequence ids are obtained.

`list_paired_OAS_studies()` returns the list of studies that populate the "Paired Sequences" database in a character vector. This list can be seen here: https://opig.stats.ox.ac.uk/webapps/ngsdb/paired/.

`list_paired_OAS_units()` returns the list of all the OAS unit files that belong to a given study from the "Paired Sequences" database.

`download_paired_OAS_units()` returns an invisible `NULL`.

`extract_metadata_from_OAS_units()` returns the metadata of all the OAS unit files found in the specified directory in a data.frame with 1 row per file.

`extract_sequences_from_paired_OAS_units()` extracts the sequence pairs from all the OAS unit files found in the specified directory and returns them in a named DNAStringSet object where the names are the sequence ids. The sequence ids are obtained by prefixing the original sequence ids found in the files with the name of the unit followed by `_heavy_chain_` or `_light_chain_`.

References

- The OAS paper:
  Tobias H. Olsen, Fergus Boyles, Charlotte M. Deane. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Science (2021). https://doi.org/10.1002/pro.4205
- The "MiAIRR format" paper:
  Rubelt, F., Busse, C., Bukhari, S. et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. Nat Immunol 18, 1274–1278 (2017). https://doi.org/10.1038/ni.3873

See Also

- OAS's homepage at: https://opig.stats.ox.ac.uk/webapps/oas/
- The `igblastn` function to run the `igblastn` *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- tibble objects implemented in the **tibble** package.
- DNAStringSet objects implemented in the **Biostrings** package.

Examples

```
list_paired_OAS_studies()

list_paired_OAS_units("Eccles_2020")

## Import all the pairs of antibody sequences from the Eccles_2020 study:
```

```
download_dir <- tempdir()
download_paired_OAS_units("Eccles_2020", destdir=download_dir)

metadata <- extract_metadata_from_OAS_units(download_dir)
metadata  # data.frame with 1 row per unit file

sequences <- extract_sequences_from_paired_OAS_units(download_dir)
sequences  # DNAStringSet object

## Odd indices correspond to heavy chain sequences and even indices
## to light chain sequences:

head(names(sequences))

sequences[1:2]  # 1st pair
sequences[3:4]  # 2nd pair
sequences[5:6]  # 3rd pair
# etc...
```

`outfmt7-utils` *Handle igblastn output format 7*

Description

Some utilities to handle igblastn output format 7.

Usage

```
list_outfmt7_specifiers()

parse_outfmt7(out_lines)
```

Arguments

`out_lines`	The character vector returned by `igblatsn(query, outfmt=7, parse.out=FALSE, ...)`.

Value

`list_outfmt7_specifiers()` returns the list of format specifiers supported by `igblastn()` formatting option 7.

`parse_outfmt7(out_lines)` returns the parsed form of `out_lines` in a list.

See Also

- The `igblastn` function to run the `igblastn` *standalone executable* included in IgBLAST from R. This is the main function in the **igblastr** package.
- IgBLAST is described at https://pubmed.ncbi.nlm.nih.gov/23671333/.

Examples

```
if (!has_igblast()) install_igblast()

## Files 'heavy_sequences.fasta' and 'light_sequences.fasta' included
## in igblastr contain 250 paired heavy- and light- chain sequences (125
## sequences in each file) downloaded from OAS (the Observed Antibody
## Space database):
filenames <- paste0(c("heavy", "light"), "_sequences.fasta")
query <- system.file(package="igblastr", "extdata", filenames)

## Keep only the first 10 sequences from each file:
query <- c(head(readDNAStringSet(query[[1L]]), n=10),
           head(readDNAStringSet(query[[2L]]), n=10))

## Select the germline and C-region dbs to use with igblastn():
use_germline_db("_AIRR.human.IGH+IGK+IGL.202501")
use_c_region_db("_IMGT.human.IGH+IGK+IGL.202412")

## ------------------------------------------------------------------------
## FIRST igblastn RUN: GET OUTPUT IN FORMAT 7
## ------------------------------------------------------------------------

## For this first run we specify 'outfmt=7' and 'parse.out=FALSE':
out_lines <- igblastn(query, outfmt=7, parse.out=FALSE)
out_lines  # raw output

out <- parse_outfmt7(out_lines)  # parse the output

## Output contains one record per query sequence:
length(out$records)  # 20

## Each record can have 5 or 6 sections:
##   1. query_details
##   2. VDJ_rearrangement_summary
##   3. VDJ_junction_details
##   4. subregion_sequence_details (can be missing)
##   5. alignment_summary
##   6. hit_table

## Taking a close look at the first record:
rec1 <- out$records[[1]]
rec1

qseqid(rec1)    # query sequence id associated with this record

rec1$hit_table  # data.frame with the standard columns

## ------------------------------------------------------------------------
## SECOND igblastn RUN: GET OUTPUT IN CUSTOMIZED FORMAT 7
## ------------------------------------------------------------------------

## For this second run we request a customized format 7 by supplying
## space delimited format specifiers:
outfmt <- "7 qseqid sseqid pident nident length score"
out <- igblastn(query, outfmt=outfmt)
```

```
## Taking a close look at the first record:
rec1 <- out$records[[1]]
rec1$hit_table  # data.frame with the requested columns (+ the
                # automatic "chaintype" column)
```

Index