Package ‘EDIRquery’

January 17, 2024

**Title**  Query the EDIR Database For Specific Gene

**Version**  1.2.0

**Description**

EDIRquery provides a tool to search for genes of interest within the Exome Database of Interspersed Repeats (EDIR). A gene name is a required input, and users can additionally specify repeat sequence lengths, minimum and maximum distance between sequences, and whether to allow a 1-bp mismatch. Outputs include a summary of results by repeat length, as well as a dataframe of query results. Example data provided includes a subset of the data for the gene GAA (ENSG00000171298). To query the full database requires providing a path to the downloaded database files as a parameter.

**License**  GPL-3

**Encoding**  UTF-8

**biocViews**  Genetics, SequenceMatching

**Roxygen**  list(markdown = TRUE)

**RoxygenNote**  7.2.3

**Depends**  R (>= 4.2.0)

**Imports**  tibble (>= 3.1.6), tictoc (>= 1.0.1), utils (>= 4.1.3), stats (>= 4.1.3), readr (>= 2.1.2), InteractionSet (>= 1.22.0), GenomicRanges (>= 1.46.1)

**Suggests**  knitr, rmarkdown, testthat (>= 3.0.0)

**VignetteBuilder**  knitr

**Config/testthat/edition**  3

**git_url**  https://git.bioconductor.org/packages/EDIRquery

**git_branch**  RELEASE_3_18

**git_last_commit**  189026a

**git_last_commit_date**  2023-10-24

**Repository**  Bioconductor 3.18

**Date/Publication**  2024-01-17

**Author**  Laura D.T. Vo Ngoc [aut, cre] (<https://orcid.org/0000-0002-1597-900X>)

**Maintainer**  Laura D.T. Vo Ngoc <doan.vongoc@vub.be>


\textbf{R topics documented:}

\begin{verbatim}
gene_chr . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .  2
gene_lookup . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .  2
\end{verbatim}

\textbf{Description}

A dataset containing two formats of gene names and associated chromosome number.

\textbf{Usage}

\begin{verbatim}
gene_chr
\end{verbatim}

\textbf{Format}

A data frame with 60571 rows and 3 variables:

- \texttt{chromosome_name} chromosome
- \texttt{ensembl_gene_id} Ensembl gene ID of gene
- \texttt{hgnc_symbol} HGNC symbol of gene

\begin{verbatim}
gene_lookup . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .  2
\end{verbatim}

\textbf{Description}

This function searches for a specified gene in the EDIR dataset. A gene name is a required parameter.

\textbf{Usage}

\begin{verbatim}
gene_lookup(  
gene,  
length = NA,  
mindist = 0,  
maxdist = 1000,  
format = "data.frame",  
summary = FALSE,  
mismatch = TRUE,  
path = NA  
)
\end{verbatim}
Arguments

- **gene**: The gene name (ENSEMBL ID or HGNC symbol)
- **length**: Repeat sequence length, must be between 7 and 20. Defaults to NA. If NA, results will include all available lengths in dataset for queried gene.
- **mindist**: Minimum spacer distance between repeats. Defaults to 0.
- **maxdist**: Maximum spacer distance between repeats. Defaults to 1000.
- **format**: Output table format. One of 'data.frame', 'GInteractions'. Defaults to 'data.frame'.
- **summary**: Logical value indicating whether to store summary. Defaults to FALSE.
- **mismatch**: Logical value indicating whether to allow 1 mismatch in sequences. Defaults to TRUE.
- **path**: String containing path to directory holding downloaded dataset files. Defaults to NA. If not provided (path = NA), gene_lookup() will use subset of data provided as example.

Details

Summary of results printed to console includes gene name, gene length (bp), Ensembl transcript ID, queried distance between repeats (default: 0-1000 bp), and an overview of total results for the given repeat length. Console outputs include runtime.

Value

A data.frame of the results from the EDIR database. If summary = TRUE, returns a tibble containing the summary ($summary), and query results ($results).

Examples

```r
## With given repeat length,
gene_lookup("GAA", length = 7, mindist = 10, maxdist = 1000,
mismatch = TRUE)

## Without specified repeat length
gene_lookup("GAA", mindist = 0, maxdist = 1000, mismatch = TRUE)

## To access query results, store in variable
output <- gene_lookup("GAA", length = 7, mindist = 10, maxdist = 1000,
mismatch = FALSE)
head(output)

## With summary = TRUE
output <- gene_lookup("GAA", length = 10, mindist = 10, maxdist = 1000,
summary = TRUE,
mismatch = TRUE)
output$summary
head(output$results)
```
Index

* datasets
  gene_chr, 2

gene_chr, 2
gene_lookup, 2