EDIRquery

Laura DT Vo Ngoc

2022-08-29

Introduction

Intragenic exonic deletions are known to contribute to genetic diseases and are often flanked by regions of homology. The Exome Database of Interspersed Repeats (EDIR) was developed to provide an overview of the positions of repetitive structures within the human genome composed of interspersed repeats encompassing a coding sequence. The package `EDIRquery` provides user-friendly tools to query this database for genes of interest.

Dataset

EDIR provides a dataset of pairwise repeat structures in which both sequences are located within a maximum of 1000 bp from each other, and fulfill one of the following selection criteria:

- >= 1 repeat located in an exon
- Both repeats situated in different introns flanking one or more exons

A subset of EDIR is provided as example data, representing a subset of the interspersed repeats data for the gene GAA (ENSG00000171298) on chromosome 17.

To query the full the database, provide the data directory to `gene_lookup()` in the `path` parameter.

Usage

```
library(EDIRquery)
```

EDIR can easily be queried using the `gene_lookup` function, using the gene name and additional parameters:

Argument	Description	Default
gene	**required:** The gene name (ENSEMBLE ID or HGNC symbol)	-
length	Repeat sequence length, must be between 7 and 20. If NA, results will include all available lengths in dataset for queried gene	NA
mindist	Minimum spacer distance (bp) between repeats	0
maxdist	Maximum spacer distance (bp) between repeats	1000
summary	Logical value indicating whether to store summary	FALSE

Argument	Description	Default
mismatch	Logical value indicating whether to allow 1 mismatch in sequence	TRUE
path	String containing path to directory holding downloaded dataset files. If not provided (`path` = NA), provided example subset of data will be used	NA

Examples

A summary of the input printed to console, including the 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.

Example querying the gene "GAA" with repeats of length 7, and allowing for 1 mismatch:

```
# Summary of results (printed to console)
gene_lookup("GAA", length = 7, mismatch = TRUE)
#>  Parameters
#>  Repeat length:   7  bp
#>  Gene:            ENSG00000171298 / GAA
#>  Gene length:     18325 bp
#>  Transcript ID:   ENST00000302262
#>  Distance:        0-1000 bp
#>  Mismatch:        TRUE
#>
#>   repeat_length unique_seqs tot_instances tot_structures avg_dist
#> 1             7        5172         10460          14562 486.2603
#>   norm_instances_bp norm_instances_Mb norm_structures_bp norm_structures_Mb
#> 1         0.5708049          570804.9          0.7946521           794652.1
#>
#> Runtime: 0.64 sec elapsed
```

If no `length` is provided, a summary of all available repeat length results will be printed:

```
# Summary of results (printed to console)
gene_lookup("GAA", mismatch = TRUE)
#>  Parameters
#>
#>  Gene:            ENSG00000171298 / GAA
#>  Gene length:     18325 bp
#>  Transcript ID:   ENST00000302262
#>  Distance:        0-1000 bp
#>  Mismatch:        TRUE
#>
#>    repeat_length unique_seqs tot_instances tot_structures avg_dist
#> 1              7        5172         10460          14562 486.2603
#> 2              8        5677          7592           7062 516.1827
#> 3              9        3160          3461           2226 508.7588
#> 4             10        1172          1227            690 500.2217
#> 5             11         389           399            209 492.5263
#> 6             12         122           124             63 454.6190
#> 7             13          42            42             21 346.2857
```

```
#> 8             14          14            14              7 271.1429
#> 9             15           4             4              2  43.0000
#> 10            16           2             2              1  42.0000
#>    norm_instances_bp norm_instances_Mb norm_structures_bp norm_structures_Mb
#> 1       0.5708049113       570804.9113       7.946521e-01       794652.11460
#> 2       0.4142974079       414297.4079       3.853752e-01       385375.17053
#> 3       0.1888676671       188867.6671       1.214734e-01       121473.39700
#> 4       0.0669577080        66957.7080       3.765348e-02        37653.47885
#> 5       0.0217735334        21773.5334       1.140518e-02        11405.18417
#> 6       0.0067667121         6766.7121       3.437926e-03         3437.92633
#> 7       0.0022919509         2291.9509       1.145975e-03         1145.97544
#> 8       0.0007639836          763.9836       3.819918e-04          381.99181
#> 9       0.0002182810          218.2810       1.091405e-04          109.14052
#> 10      0.0001091405          109.1405       5.457026e-05           54.57026
#>
#> Runtime: 0.64 sec elapsed
```

Storing the output in a variable allows viewing of the individual results in the output dataframe:

```
# Database output of query
results <- gene_lookup("GAA", length = 7, mismatch = TRUE)
#>  Parameters
#>  Repeat length:   7  bp
#>  Gene:            ENSG00000171298 / GAA
#>  Gene length:     18325 bp
#>  Transcript ID:   ENST00000302262
#>  Distance:        0-1000 bp
#>  Mismatch:        TRUE
#>
#>   repeat_length unique_seqs tot_instances tot_structures avg_dist
#> 1             7        5172         10460          14562 486.2603
#>   norm_instances_bp norm_instances_Mb norm_structures_bp norm_structures_Mb
#> 1         0.5708049          570804.9          0.7946521           794652.1
#>
#> Runtime: 0.33 sec elapsed
head(results)
#>      chromosome repeat_length repeat_seq    start      end repeat_seq2   start2
#> 3930         17             7    CCGCGGG 80101595 80101602     CCGCGGG 80101734
#> 3931         17             7    CCGAGGC 80105602 80105609     CCGAGGA 80105843
#> 3932         17             7    CGGAGGG 80110005 80110012     GCGAGGG 80110061
#> 3933         17             7    CCAAGGG 80118254 80118261     CCGAGGG 80118270
#> 3934         17             7    CCGAGGG 80118270 80118277     GCGAGGG 80118318
#> 3935         17             7    CCGAGGG 80118270 80118277     CAGAGGG 80118533
#>          end2 distance ensembl_gene_id hgnc_symbol        gene_range
#> 3930 80101741      132 ENSG00000171298         GAA 80101556-80119881
#> 3931 80105850      234 ENSG00000171298         GAA 80101556-80119881
#> 3932 80110068       49 ENSG00000171298         GAA 80101556-80119881
#> 3933 80118277        9 ENSG00000171298         GAA 80101556-80119881
#> 3934 80118325       41 ENSG00000171298         GAA 80101556-80119881
#> 3935 80118540      256 ENSG00000171298         GAA 80101556-80119881
#>      ensembl_transcript_id  transcript_range intron_exon intron_exon2
#> 3930       ENST00000302262 80101581-80101890          E1           E1
#> 3931       ENST00000302262 80105133-80105748          I2           E3
```

```
#> 3932       ENST00000302262 80109945-80110055          E9          I9
#> 3933       ENST00000302262 80118193-80118357         E18         E18
#> 3934       ENST00000302262 80118193-80118357         E18         E18
#> 3935       ENST00000302262 80118193-80118357         E18         I18
#>                  feature mismatch
#> 3930            same exon        0
#> 3931 spanning intron-exon        1
#> 3932 spanning intron-exon        1
#> 3933            same exon        1
#> 3934            same exon        1
#> 3935 spanning intron-exon        1
```

Session info

```
# Database output of query
sessionInfo()
#> R version 4.1.3 (2022-03-10)
#> Platform: x86_64-w64-mingw32/x64 (64-bit)
#> Running under: Windows 10 x64 (build 19043)
#>
#> Matrix products: default
#>
#> locale:
#> [1] LC_COLLATE=English_Belgium.1252  LC_CTYPE=English_Belgium.1252
#> [3] LC_MONETARY=English_Belgium.1252 LC_NUMERIC=C
#> [5] LC_TIME=English_Belgium.1252
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#> [1] EDIRquery_0.99.0
#>
#> loaded via a namespace (and not attached):
#>  [1] rstudioapi_0.13  knitr_1.38       magrittr_2.0.3   hms_1.1.1
#>  [5] tidyselect_1.1.2 bit_4.0.4        R6_2.5.1         rlang_1.0.3
#>  [9] fastmap_1.1.0    fansi_1.0.3      stringr_1.4.0    tools_4.1.3
#> [13] tictoc_1.0.1     vroom_1.5.7      xfun_0.30        utf8_1.2.2
#> [17] cli_3.3.0        htmltools_0.5.2  ellipsis_0.3.2   bit64_4.0.5
#> [21] yaml_2.3.5       digest_0.6.29    tibble_3.1.7     lifecycle_1.0.1
#> [25] crayon_1.5.1     purrr_0.3.4      readr_2.1.2      tzdb_0.3.0
#> [29] vctrs_0.4.1      glue_1.6.2       evaluate_0.15    rmarkdown_2.14
#> [33] stringi_1.7.6    compiler_4.1.3   pillar_1.8.0     pkgconfig_2.0.3
```