

EDIRquery

Laura D.T. Vo Ngoc

2024-05-15

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

Installation

To install this package, enter the following in R:

```
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("EDIRquery")
```

Then load the package:

```
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)	-

Argument	Description	Default
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
format	Output table format. One of 'data.frame', 'GInteractions'.	'data.frame'
summary	Logical value indicating whether to store summary	FALSE
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 (<code>path = NA</code>), 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: 1.03 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: 1.14 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: 0.85 sec elapsed

head(results)
#>   chromosome repeat_length start1 end1 start2 end2 repeat_seq1
#> 3930         17          7 80101595 80101601 80101734 80101740   CCGCGGG
#> 3931         17          7 80105602 80105608 80105843 80105849   CCGAGGC

```

```

#> 3932      17      7 80110005 80110011 80110061 80110067      CCGAGGG
#> 3933      17      7 80118254 80118260 80118270 80118276      CCAAGGG
#> 3934      17      7 80118270 80118276 80118318 80118324      CCGAGGG
#> 3935      17      7 80118270 80118276 80118533 80118539      CCGAGGG
#>      intron_exon1 repeat_seq2 intron_exon2 distance ensembl_gene_id hgnc_symbol
#> 3930      E1      CCGCGGG      E1      132 ENSG00000171298      GAA
#> 3931      I2      CCGAGGA      E3      234 ENSG00000171298      GAA
#> 3932      E9      GCGAGGG      I9      49  ENSG00000171298      GAA
#> 3933      E18     CCGAGGG      E18      9  ENSG00000171298      GAA
#> 3934      E18     GCGAGGG      E18      41 ENSG00000171298      GAA
#> 3935      E18     CAGAGGG      I18     256 ENSG00000171298      GAA
#>      gene_range ensembl_transcript_id transcript_range
#> 3930 80101556-80119881      ENST00000302262 80101581-80101890
#> 3931 80101556-80119881      ENST00000302262 80105133-80105748
#> 3932 80101556-80119881      ENST00000302262 80109945-80110055
#> 3933 80101556-80119881      ENST00000302262 80118193-80118357
#> 3934 80101556-80119881      ENST00000302262 80118193-80118357
#> 3935 80101556-80119881      ENST00000302262 80118193-80118357
#>      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

```

Changing the format parameter to `GInteractions` returns a `GenomicInteractions` object instead of a dataframe:

```

# Database output of query
results <- gene_lookup("GAA", length = 7, format = "GInteractions", 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: 1 sec elapsed

head(results)
#> GInteractions object with 6 interactions and 11 metadata columns:
#>      seqnames1      ranges1      seqnames2      ranges2 /
#>      <Rle>      <IRanges>      <Rle>      <IRanges> /
#> [1]      17 80101595-80101601 ---      17 80101734-80101740 /

```

```

#> [2] 17 80105602-80105608 --- 17 80105843-80105849 /
#> [3] 17 80110005-80110011 --- 17 80110061-80110067 /
#> [4] 17 80118254-80118260 --- 17 80118270-80118276 /
#> [5] 17 80118270-80118276 --- 17 80118318-80118324 /
#> [6] 17 80118270-80118276 --- 17 80118533-80118539 /
#> anchor1.repeat_seq1 anchor1.intron_exon1 anchor2.repeat_seq2
#> <character> <character> <character>
#> [1] CCGCGGG E1 CCGCGGG
#> [2] CCGAGGC I2 CCGAGGA
#> [3] CGGAGGG E9 GCGAGGG
#> [4] CCAAGGG E18 CCGAGGG
#> [5] CCGAGGG E18 GCGAGGG
#> [6] CCGAGGG E18 CAGAGGG
#> anchor2.intron_exon2 ensembl_gene_id hgnc_symbol gene_range
#> <character> <character> <character> <character>
#> [1] E1 ENSG00000171298 GAA 80101556-80119881
#> [2] E3 ENSG00000171298 GAA 80101556-80119881
#> [3] I9 ENSG00000171298 GAA 80101556-80119881
#> [4] E18 ENSG00000171298 GAA 80101556-80119881
#> [5] E18 ENSG00000171298 GAA 80101556-80119881
#> [6] I18 ENSG00000171298 GAA 80101556-80119881
#> ensembl_transcript_id transcript_range feature mismatch
#> <character> <character> <character> <integer>
#> [1] ENST00000302262 80101581-80101890 same exon 0
#> [2] ENST00000302262 80105133-80105748 spanning intron-exon 1
#> [3] ENST00000302262 80109945-80110055 spanning intron-exon 1
#> [4] ENST00000302262 80118193-80118357 same exon 1
#> [5] ENST00000302262 80118193-80118357 same exon 1
#> [6] ENST00000302262 80118193-80118357 spanning intron-exon 1
#> -----
#> regions: 10460 ranges and 0 metadata columns
#> seqinfo: 1 sequence from an unspecified genome; no seqlengths

```

Session info

```

# Database output of query
sessionInfo()
#> R version 4.4.0 RC (2024-04-16 r86468 ucrt)
#> Platform: x86_64-w64-mingw32/x64
#> Running under: Windows Server 2022 x64 (build 20348)
#>
#> Matrix products: default
#>
#> locale:
#> [1] LC_COLLATE=C
#> [2] LC_CTYPE=English_United States.utf8
#> [3] LC_MONETARY=English_United States.utf8
#> [4] LC_NUMERIC=C
#> [5] LC_TIME=English_United States.utf8
#>
#> time zone: America/New_York

```

```

#> tzcode source: internal
#>
#> attached base packages:
#> [1] stats      graphics  grDevices  utils      datasets  methods    base
#>
#> other attached packages:
#> [1] EDIRquery_1.5.0
#>
#> loaded via a namespace (and not attached):
#> [1] bit_4.0.5                Matrix_1.7-0
#> [3] jsonlite_1.8.8           compiler_4.4.0
#> [5] crayon_1.5.2             tidyselect_1.2.1
#> [7] Rcpp_1.0.12              SummarizedExperiment_1.35.0
#> [9] Biobase_2.65.0           GenomicRanges_1.57.0
#> [11] IRanges_2.39.0          yaml_2.3.8
#> [13] fastmap_1.2.0           lattice_0.22-6
#> [15] readr_2.1.5             R6_2.5.1
#> [17] XVector_0.45.0          S4Arrays_1.5.0
#> [19] GenomeInfoDb_1.41.0     knitr_1.46
#> [21] BiocGenerics_0.51.0     tibble_3.2.1
#> [23] DelayedArray_0.31.1     MatrixGenerics_1.17.0
#> [25] GenomeInfoDbData_1.2.12 tzdb_0.4.0
#> [27] pillar_1.9.0           rlang_1.1.3
#> [29] utf8_1.2.4             xfun_0.44
#> [31] bit64_4.0.5            SparseArray_1.5.4
#> [33] cli_3.6.2              magrittr_2.0.3
#> [35] tictoc_1.2.1           zlibbioc_1.51.0
#> [37] digest_0.6.35          InteractionSet_1.33.0
#> [39] grid_4.4.0             vroom_1.6.5
#> [41] hms_1.1.3              lifecycle_1.0.4
#> [43] S4Vectors_0.43.0       vctrs_0.6.5
#> [45] glue_1.7.0             evaluate_0.23
#> [47] abind_1.4-5            stats4_4.4.0
#> [49] fansi_1.0.6            rmarkdown_2.26
#> [51] httr_1.4.7             pkgconfig_2.0.3
#> [53] matrixStats_1.3.0      tools_4.4.0
#> [55] htmltools_0.5.8.1      UCSC.utils_1.1.0

```