Package 'scoper'

August 5, 2019

Type Package

Version 0.2.0

Date 2019-08-05

Title Spectral Clustering-Based Method for Identifying B Cell Clones

Description Provides a computational framework for B cell clones identification
from adaptive immune receptor repertoire sequencing (AIRR-Seq) datasets.
Three models are included (identical, hierarchical, and spectral)
that perform clustering among sequences of BCRs/IGs (B cell receptors/immunoglobulins)
which share the same V gene, J gene and junction length.
Nouri N and Kleinstein SH (2018) <doi: 10.1093/bioinformatics/bty235>.
Gupta NT, et al. (2017) <doi: 10.4049/jimmunol.1601850>.

License CC BY-SA 4.0

URL https://scoper.readthedocs.io

BugReports https://bitbucket.org/kleinstein/scoper/issues

LazyData true

BuildVignettes true

VignetteBuilder knitr

Encoding UTF-8

SystemRequirements C++11

Depends R (>= 3.1.2), ggplot2 (>= 3.2.0)

Imports alakazam (>= 0.3.0), shazam (>= 0.2.0), doParallel, foreach,
dplyr (>= 0.8.1), Rcpp (>= 0.12.12), seqinr, data.table,
stringi, stringr, methods, stats, rlang

LinkingTo Rcpp

Suggests knitr, rmarkdown, testthat

RoxygenNote 6.1.1

Collate 'Data.R' 'Scoper.R' 'Functions.R' 'RcppExports.R'

NeedsCompilation yes

Author Nima Nouri [aut, cre],
Jason Vander Heiden [ctb],
Steven Kleinstein [aut, cph]

Maintainer Nima Nouri <nima.nouri@yale.edu>

Repository CRAN

Date/Publication 2019-08-05 21:50:02 UTC

R topics documented:

`defineClonesScoper` *Assigning Ig sequences into clonal groups*

Description

The `defineClonesScoper` function provides a computational pipline for assigning Ig sequences into clonal groups sharing same V gene, J gene, and junction length.

Usage

```
defineClonesScoper(db, model = c("identical", "hierarchical",
  "spectral"), method = c("nt", "aa", "single", "average", "complete",
  "novj", "vj"), germline_col = "GERMLINE_IMGT",
  sequence_col = "SEQUENCE_IMGT", junction_col = "JUNCTION",
  v_call_col = "V_CALL", j_call_col = "J_CALL",
  clone_col = c("clone_id", "CLONE"), targeting_model = NULL,
  len_limit = NULL, first = FALSE, cdr3 = FALSE, mod3 = FALSE,
  max_n = NULL, threshold = NULL, base_sim = 0.95, iter_max = 1000,
  nstart = 1000, nproc = 1, verbose = FALSE, log_verbose = FALSE,
  out_dir = ".", summerize_clones = FALSE)
```

Arguments

`db`	data.frame containing sequence data.
`model`	one of the `"identical"`, `"hierarchical"`, or `"spectral"`. See Details for description.
`method`	one of the `"nt"`, `"aa"`, `"single"`, `"average"`, `"complete"`, `"novj"`, or `"vj"`. See Details for description.
`germline_col`	character name of the column containing the germline or reference sequence.
`sequence_col`	character name of the column containing input sequences.

junction_col — character name of the column containing junction sequences. Also used to determine sequence length for grouping.

v_call_col — character name of the column containing the V-segment allele calls.

j_call_col — character name of the column containing the J-segment allele calls.

clone_col — one of the "CLONE" or "clone_id" for the output column name containing the clone ids.

targeting_model — TargetingModel object. Only applicable if model = "spectral" and method = "vj". See Details for description.

len_limit — IMGT_V object defining the regions and boundaries of the Ig sequences. If NULL, mutations are counted for entire sequence. Only applicable if model = "spectral" and method = "vj".

first — specifies how to handle multiple V(D)J assignments for initial grouping. If TRUE only the first call of the gene assignments is used. If FALSE the union of ambiguous gene assignments is used to group all sequences with any overlapping gene calls.

cdr3 — if TRUE removes 3 nts from both ends of "junction_col" (converts IMGT junction to CDR3 region). if TRUE remove junction_col(s) with length less than 7 nts.

mod3 — if TRUE removes junction_col(s) with number of nucleotides not modulus of 3.

max_n — The maximum number of N's to permit in the junction sequence before excluding the record from clonal assignment. Note, under model "hierarchical" and method "single" non-informative positions can create artifactual links between unrelated sequences. Use with caution. Default is set to be "NULL" for no action.

threshold — the distance threshold for clonal grouping if model = "hierarchical"; or the upper-limit cut-off if model = "spectral".

base_sim — required similarity cut-off for sequences in equal distances from each other. Only applicable if model = "spectral".

iter_max — the maximum number of iterations allowed for kmean clustering step.

nstart — the number of random sets chosen for kmean clustering initialization.

nproc — number of cores to distribute the function over.

verbose — if TRUE report a summary of each step cloning process; if FALSE process cloning silently.

log_verbose — if TRUE write verbose logging to a file in out_dir.

out_dir — specify the output directory to save log_verbose. The input file directory is used if this is not specified.

summerize_clones — if TRUE performs a series of analysis to assess the clonal landscape. See Value for description.

Details

`defineClonesScoper` provides a computational platform to explore the B cell clonal relationships in high-throughput Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) data sets. Three models are included which perform clustering among sequences of B cell receptors (BCRs, also referred to as Immunoglobulins, (Igs)) that share the same V gene, J gene and junction length:

- `model = "identical"`: defines clones among identical junctions. Available `method`(s) are: (1) `"nt"` (nucleotide based clustering) and (2) `"aa"` (amino acid based clustering).
- `model = "hierarchical"`: hierarchical clustering-based method for partitioning sequences into clones. Availabe agglomeration `method`(s) are: (1) `"single"`, (2) `"average"`, and (3) `"complete"`. The fixed `threshold` (a numeric scalar where the tree should be cut) must be provided.
- `model = "spectral"`: provides an unsupervised pipline for assigning Ig sequences into clonal groups. If `method = "novj"`, clonal relationships are inferred using an adaptive threshold that indicates the level of similarity among junction sequences in a local neighborhood. If `method = "vj"`: clonal relationships are inferred not only based on the junction region homology, but also takes into account the mutation profiles in the V and J segments. `germline_col` and `sequence_col` must be provided. Mutation counts are determined by comparing the input sequences (in the column specified by `sequence_col`) to the effective germline sequence (calculated from sequences in the column specified by `germline_col`). Not mandatory, but the influence of SHM hot- and cold-spot biases in the clonal inference process will be noted if a SHM targeting model is provided through argument `targeting_model` (see createTargetingModel for more technical details).

Value

For `summerize_clones = FALSE`, a modified data.frame with clone identifiers in the `clone_col` column. For `summerize_clones = TRUE` returns a list containing:

- `db`: modified `db` data.frame with clone identifiers in the `clone_col` column.
- `vjl_group_summ`: data.frame of clones summary, e.g. size, V-gene, J-gene, junction lentgh, and so on.
- `inter_intra`: data.frame containing minimum inter (between) and maximum intra (within) clonal distances.
- `eff_threshold`: effective cut-off separating the inter (between) and intra (within) clonal distances.
- `plot_inter_intra`: ggplot histogram of inter (between) versus intra (within) clonal distances. The effective threshold is shown with a horizental dashed-line.

If `log_verbose = TRUE`, it will write verbose logging to a file in the current directory or the specified `out_dir`.

Examples

```
results <- defineClonesScoper(ExampleDb,
                              model="hierarchical", method="single",
                              threshold=0.15, summerize_clones=TRUE)
```

ExampleDb *Example Change-O database*

Description

A small example database subset from Laserson and Vigneault et al, 2014.

Usage

```
ExampleDb
```

Format

A data.frame with the following Change-O style columns:

- SEQUENCE_ID: Sequence identifier
- SEQUENCE_IMGT: IMGT-gapped observed sequence.
- GERMLINE_IMGT_D_MASK: IMGT-gapped germline sequence with N, P and D regions masked.
- V_CALL: V region allele assignments.
- V_CALL_GENOTYPED: TIgGER corrected V region allele assignment.
- D_CALL: D region allele assignments.
- J_CALL: J region allele assignments.
- JUNCTION: Junction region sequence.
- JUNCTION_LENGTH: Length of the junction region in nucleotides.
- NP1_LENGTH: Combined length of the N and P regions proximal to the V region.
- NP2_LENGTH: Combined length of the N and P regions proximal to the J region.
- SAMPLE: Sample identifier. Time in relation to vaccination.
- ISOTYPE: Isotype assignment.
- DUPCOUNT: Copy count (number of duplicates) of the sequence.

References

1. Laserson U and Vigneault F, et al. High-resolution antibody dynamics of vaccine-induced immune responses. Proc Natl Acad Sci USA. 2014 111:4928-33.

scoper *The SCOPer package*

Description

Provides a computational framework for B cell clones identification from adaptive immune receptor repertoire sequencing (AIRR-Seq) datasets. Three models are included (identical, hierarchical, and spectral) which perform clustering among sequences of B cell receptors (BCRs, also referred to as Immunoglobulins, (Igs)) that share the same V gene, J gene and junction length.

SCOPer

- defineClonesScoper: Clustering sequences into clonal groups.

References

1. Nouri N and Kleinstein SH (2018). A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data. Bioinformatics, 34(13):i341-i349.
2. Gupta NT, et al. (2017). Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. The Journal of Immunology, 198(6):2489-2499.

Index