Package 'IgGeneUsage'

April 12, 2022

Type Package

Title Differential gene usage in immune repertoires

Version 1.8.0

Description Detection of biases in immunoglobulin (Ig) gene usage between adaptive immune repertoires that belong to different biological conditions is an important task in immune repertoire profiling. IgGeneUsage detects aberrant Ig gene usage using probabilistic model which is analyzed computationally by Bayes inference.

License file LICENSE

Depends methods, R (>= 3.6.0), Rcpp (>= 0.12.0), SummarizedExperiment, StanHeaders (> 2.18.1)

Imports rstan (>= 2.19.2), reshape2 (>= 1.4.3)

Suggests BiocStyle, knitr, rmarkdown, testthat (>= 2.1.0), ggplot2, ggforce, gridExtra, ggrepel

Encoding UTF-8

LazyData true

NeedsCompilation no

biocViews DifferentialExpression, Regression, Genetics, Bayesian

BugReports https://github.com/snaketron/IgGeneUsage/issues

RoxygenNote 6.1.1

VignetteBuilder knitr

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

git_branch RELEASE_3_14

git_last_commit b71fcc0

git_last_commit_date 2021-10-26

Date/Publication 2022-04-12

Author Simo Kitanovski [aut, cre]

Maintainer Simo Kitanovski <simo.kitanovski@uni-due.de>

R topics documented:

`CDR3_Epitopes` *Net charge usage in CDR3 sequences of T-cell receptor repertoires disturbed by Influenza-A and CMV*

Description

Data of CDR3 sequence from human T-cells receptors (TRB-chain) downloaded from VDJdb. CDR3 sequences annotated to epitopes in Influenza-A and CMV were selected from different publications, as long as the publication contains at least 100 CDR3 sequences. Each publication is considered as a repertoire (sample).

To compute the net CDR3 sequence charge, we consider the amino acids K, R and H as +1 charged, while D and E as -1 charged. Thus, we computed the net charge of a CDR3 sequence by adding up the individual residue charges.

Usage

```
data("CDR3_Epitopes")
```

Format

A data frame with 4 columns: "sample_id", "condition", "gene_name" and "gene_usage_count". The format of the data is suitible to be used as input in IgGeneUsage

gene_name = net charge group

Source

https://vdjdb.cdr3.net/

Examples

```
data(CDR3_Epitopes)
head(CDR3_Epitopes)
```

DGU *Differential gene usage in immune repertoires*

Description

IgGeneUsage detects differential gene usage in immune repertoires that belong to two biological conditions.

Usage

```
DGU(usage.data, mcmc.warmup, mcmc.steps,
    mcmc.chains, mcmc.cores, hdi.level,
    adapt.delta, max.treedepth)
```

Arguments

`usage.data`	Data.frame with 4 columns: 'sample_id' = character identifier of each repertoire, 'condition' = character key representing each of the two biological conditions, 'gene_name' = character name of each gene to be tested for differential usage, 'gene_usage_count' = number of rearrangements belonging to a specific sample_id x condition x gene_name. Alternatively, usage.data can be a SummarizedExperiment object. See examplary data 'data(Ig_SE)' for more information.
`mcmc.chains, mcmc.warmup, mcmc.steps, mcmc.cores`	Number of MCMC chains (default = 4), number of cores to use (default = 1), length of MCMC chains (default = 1,500), length of adaptive part of MCMC chains (default = 500).
`hdi.level`	Highest density interval (HDI) (default = 0.95).
`adapt.delta`	MCMC setting (default = 0.95).
`max.treedepth`	MCMC setting (default = 12).

Details

The input to IgGeneUsage is a table with usage frequencies for each gene of a repertoire that belongs to a particular biological condition. For the analysis of differential gene usage between two biological conditions, IgGeneUsage employs a Bayesian hierarchical model for zero-inflated beta-binomial (ZIBB) regression (see vignette 'User Manual: IgGeneUsage').

Value

`glm.summary`	differential gene usage statistics for each gene. 1) es = effect size on differential gene usage (mean, median standard error (se), standard deviation (sd), L (low boundary of HDI), H (high boundary of HDI); 2) contrast = direction of the effect; 3) pmax = probability of differential gene usage

test.summary	differential gene usage statistics computed with the Welch's t-test (columns start with 't'), and Wilcoxon signed-rank test (columns start with 'u'). For both test report P-values, FDR-corrected P-values, Bonferroni-corrected P-values. Additionally, we report t-value and 95% CI (from the t-test) and U-value (from the Wilcoxon signed-rank test).
glm	stanfit object
ppc.data	two types of posterior predictive checks: 1) repertoire- specific, 2) gene-specific
usage.data	processed gene usage data used for the model

Author(s)

Simo Kitanovski <simo.kitanovski@uni-due.de>

See Also

LOO, Ig, IGHV_Epitopes, IGHV_HCV, Ig_SE

Examples

```
# input data
# data(Ig)
# head(Ig)

# Alternative:
# use SummarizedExperiment input data
# data(Ig_SE)


# run differential gene usage (DGU)
# M <- DGU(usage.data = Ig,
#          mcmc.warmup = 250,
#          mcmc.steps = 1000,
#          mcmc.chains = 2,
#          mcmc.cores = 1,
#          hdi.level = 0.95,
#          adapt.delta = 0.95,
#          max.treedepth = 13)
```

Ig *IGHV gene family usage in vaccine-challenged B-cell repertoires*

Description

A small example database subset from study evaluating vaccine-induced changes in B-cell populations publicly provided by R-package alakazam (version 0.2.11). It contains IGHV gene family usage, reported in four B-cell populations (samples IgM, IgD, IgG and IgA) across two timepoints (conditions = -1 hour and +7 days).

Usage

```
data("Ig")
```

Format

A data frame with 4 columns: "sample_id", "condition", "gene_name" and "gene_usage_count". The format of the data is suitible to be used as input in IgGeneUsage

Source

R package: alakazam version 0.2.11

References

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.

Examples

```
data(Ig)
head(Ig)
```

IGHV_HCV *IGHV gene usage in HCV+ and healthy individuals*

Description

Publicly available dataset of IGHV segment usage in memory B-cells of 22 HCV+ individuals and 7 healthy donors.

Usage

```
data("IGHV_HCV")
```

Format

A data frame with 4 columns: "sample_id", "condition", "gene_name" and "gene_usage_count". The format of the data is suitible to be used as input in IgGeneUsage

Source

Tucci, Felicia A., et al. "Biased IGH VDJ gene repertoire and clonal expansions in B cells of chronically hepatitis C virus–infected individuals." Blood 131.5 (2018): 546-557.

Examples

```
data(IGHV_HCV)
head(IGHV_HCV)
```

Ig_SE *IGHV gene family usage in vaccine-challenged B-cell repertoires (SummarizedExperiment object)*

Description

A small example database subset from study evaluating vaccine-induced changes in B-cell populations publicly provided by R-package alakazam (version 0.2.11). It contains IGHV gene family usage, reported in four B-cell populations (samples IgM, IgD, IgG and IgA) across two timepoints (conditions = -1 hour and +7 days).

Usage

```
data("Ig_SE")
```

Format

A SummarizedExperiment object with 1) assay data (rows = gene name, columns = repertoires) and 2) column data.frame in which the sample names and the corresponding biological condition labels are noted.

Source

R package: alakazam version 0.2.11

References

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.

Examples

```
# inspect the data
data(Ig_SE)

# repertoire information: must have the two columns: 'condition', 'sample_id'
SummarizedExperiment::colData(Ig_SE)

# assay counts (gene frequency usage)
SummarizedExperiment::assay(x = Ig_SE)
```

LOO *Leave-one-out analysis for quantitative evaluation of the probability of DGU*

Description

IgGeneUsage detects differential gene usage in immune repertoires that belong to two biological conditions with its function DGU. To assert quantitatively the robustness of the estimated probability of DGU (pi), IgGeneUsage has a built-in procedure for a fully Bayesian leave-one-out (LOO) analysis. During each step of LOO, we discard the data of one of the repertoires, and use the remaining data to analyze for DGU with IgGeneUsage. In each step we recorded pi for all genes. Therefore, by evaluating the variability of pi for a given gene, we can we assert quantitatively its robustness.

Notice, however, that for datasets that include many repertoires (e.g. 100) LOO can be computationally costly.

Usage

```
LOO(usage.data, mcmc.warmup, mcmc.steps,
    mcmc.chains, mcmc.cores, hdi.level,
    adapt.delta, max.treedepth)
```

Arguments

`usage.data`	Data.frame with 4 columns: 'sample_id' = character identifier of each repertoire, 'condition' = character key representing each of the two biological conditions, 'gene_name' = character name of each gene to be tested for differential usage, 'gene_usage_count' = number of rearrangements belonging to a specific sample_id x condition x gene_name. Alternatively, usage.data can be a SummarizedExperiment object. See examplary data 'data(Ig_SE)' for more information.
`mcmc.chains, mcmc.warmup, mcmc.steps, mcmc.cores`	Number of MCMC chains (default = 4), number of cores to use (default = 1), length of MCMC chains (default = 1,500), length of adaptive part of MCMC chains (default = 500).
`hdi.level`	Highest density interval (HDI) (default = 0.95).
`adapt.delta`	MCMC setting (default = 0.95).
`max.treedepth`	MCMC setting (default = 12).

Details

IgGeneUsage invokes the function DGU in each LOO step. For more details see help for DGU or vignette 'User Manual: IgGeneUsage'.

Value

`loo.summary` differential gene usage statistics for each gene of a given LOO step. 1) es = effect size on differential gene usage (mean, median standard error (se), standard deviation (sd), L (low boundary of HDI), H (high boundary of HDI); 2) contrast = direction of the effect; 3) pmax = probability of differential gene usage; 4) loo.id (LOO step ID); 5 Neff (effective sample size), Rhat (potential scale reduction factor)

Author(s)

Simo Kitanovski <simo.kitanovski@uni-due.de>

See Also

DGU, Ig, IGHV_Epitopes, IGHV_HCV, Ig_SE

Examples

```
# input data:
# data(Ig)
# head(Ig)

# Alternative:
# use SummarizedExperiment input data
# data(Ig_SE)


# run leave-one-out (LOO)
# L <- LOO(usage.data = Ig,
#          mcmc.warmup = 250,
#          mcmc.steps = 1000,
#          mcmc.chains = 2,
#          mcmc.cores = 1,
#          hdi.level = 0.95,
#          adapt.delta = 0.95,
#          max.treedepth = 13)
```

Index