Package 'rabhit'

May 8, 2019

Type Package

Title Inference Tool for Antibody Haplotype

Version 0.1.1

Description Infers V-D-J haplotypes and gene deletions from AIRR-seq data,
based on IGHJ, IGHD or IGHV as anchor, by adapting a Bayesian framework.
It also calculates a Bayes factor, a number that indicates the certainty level of the inference, for each haplotyped gene.
Citation:
Gidoni, et al (2019) <doi:10.1038/s41467-019-08489-3>.

License CC BY-SA 4.0

URL https://yaarilab.bitbucket.io/RAbHIT/

BugReports https://bitbucket.org/yaarilab/haplotyper/issues

LazyData true

BuildVignettes true

VignetteBuilder knitr

Encoding UTF-8

Depends R (>= 3.2.5), ggplot2 (>= 2.0.0)

Imports dplyr (>= 0.5.0), reshape2, plotly (>= 4.7.1), gtools (>= 3.5.0), cowplot (>= 0.9.1), stats, dendextend (>= 1.9.0), data.table, ggdendro (>= 0.1.20), gridExtra, alakazam (>= 0.2.10), tigger (>= 0.2.11), methods, htmlwidgets, gtable, grDevices, rlang, RColorBrewer, tidyr, stringi

Suggests knitr, rmarkdown

RoxygenNote 6.1.1

NeedsCompilation no

Collate 'Data.R' 'rabhit.R' 'internal_functions.R' 'functions.R' 'graphic_functions.R'

Author Ayelet Peres [aut, cre],
Moriah Gidoni [aut],
Gur Yaari [aut, cph]

Maintainer Ayelet Peres <peresay@biu.ac.il>

Repository CRAN

Date/Publication 2019-05-08 13:10:03 UTC

R topics documented:

`createFullHaplotype` *Anchor gene haplotype inference*

Description

The `createFullHaplotype` functions infers haplotype based on an anchor gene.

Usage

```
createFullHaplotype(clip_db, toHap_col = c("V_CALL", "D_CALL"),
  hapBy_col = "J_CALL", hapBy = "IGHJ6", toHap_GERM,
  relative_freq_priors = TRUE, kThreshDel = 3, rmPseudo = TRUE,
  deleted_genes = c(), nonReliable_Vgenes = c(),
  min_minor_fraction = 0.3, chain = c("IGH", "IGK", "IGL"))
```

Arguments

`clip_db`	a `data.frame` in Change-O format. See details.
`toHap_col`	a vector of column names for which a haplotype should be inferred. Default is V_CALL and D_CALL.
`hapBy_col`	column name of the anchor gene. Default is J_CALL.

hapBy	a string of the anchor gene name. Default is IGHJ6.
toHap_GERM	a vector of named nucleotide germline sequences matching the allele calls in toHap_col columns in clip_db.
relative_freq_priors	if TRUE, the priors for Bayesian inference are estimated from the relative frequencies in clip_db. Else, priors are set to c(0.5,0.5). Defualt is TRUE
kThreshDel	the minimum lK (log10 of the Bayes factor) to call a deletion. Defualt is 3.
rmPseudo	if TRUE non-functional and pseudo genes are removed. Defualt is TRUE.
deleted_genes	double chromosome deletion summary table. A data.frame created by deletionsByBinom.
nonReliable_Vgenes	a list of known non reliable gene assignments. A list created by nonReliableVGenes.
min_minor_fraction	the minimum minor allele fraction to be used as an anchor gene. Default is 0.3
chain	the IG chain: IGH,IGK,IGL. Default is IGH.

Details

Function accepts a data.frame in Change-O format (https://changeo.readthedocs.io/en/version-0.4.1---airr-standards/standard.html) containing the following columns:

- 'SUBJECT': The subject name
- 'V_CALL': V allele call(s) (in an IMGT format)
- 'D_CALL': D allele call(s) (in an IMGT format, only for heavy chains)
- 'J_CALL': J allele call(s) (in an IMGT format)

Value

A data.frame, in which each row is the haplotype inference summary of a gene from the column selected in toHap_col.

The output containes the following columns:

- SUBJECT: the subject name.
- GENE: the gene name.
- Anchor gene allele 1: the haplotype inference for chromosome one. The column name is the anchor gene with the first allele.
- Anchor gene allele 2: the haplotype inference for chromosome two. The column name is the anchor gene with the second allele.
- ALLELES: allele calls for the gene.
- PRIORS_ROW: priors based on relative allele usage of the anchor gene.
- PRIORS_COL: priors based on relative allele usage of the inferred gene.
- COUNTS1: the appereance count on each chromosome of the first allele from ALLELES, the counts are seperated by a comma.
- K1: the Bayesian factor value for the first allele (from ALLELES) inference.

- COUNTS2: the appereance count on each chromosome of the second allele from ALLELES, the counts are seperated by a comma.
- K2: the Bayesian factor value for the second allele (from ALLELES) inference.
- COUNTS3: the appereance count on each chromosome of the third allele from ALLELES, the counts are seperated by a comma.
- K3: the Bayesian factor value for the third allele (from ALLELES) inference.
- COUNTS4: the appereance count on each chromosome of the fourth allele from ALLELES, the counts are seperated by a comma.
- K4: the Bayesian factor value for the fourth allele (from ALLELES) inference.

Examples

```
# Load example data and germlines
data(samples_db, HVGERM, HDGERM)

# Selecting a single individual
clip_db = samples_db[samples_db$SUBJECT=='I5', ]

# Infering haplotype
haplo_db = createFullHaplotype(clip_db,toHap_col=c('V_CALL','D_CALL'),
hapBy_col='J_CALL',hapBy='IGHJ6',toHap_GERM=c(HVGERM,HDGERM))
```

---

deletionHeatmap *Graphical output of single chromosome deletions*

Description

The deletionHeatmap function generates a graphical output of the single chromosome deletions in multiple samples.

Usage

```
deletionHeatmap(hap_table, kThreshDel = 3, chain = c("IGH", "IGK",
  "IGL"))
```

Arguments

hap_table	haplotype summary table. See details.
kThreshDel	the minimum lK (log10 of the Bayes factor) used in createFullHaplotype to call a deletion. Indicates the color for strong deletion. Defualt is 3.
chain	the IG chain: IGH,IGK,IGL. Default is IGH.

Details

A data.frame created by createFullHaplotype.

Value

A single chromosome deletion visualization.

Examples

```
# Plotting single choromosme deletion from haplotype inference
deletionHeatmap(samplesHaplotype)
```

`deletionsByBinom` *Double chromosome deletion by relative gene usage*

Description

The `deletionsByBinom` function inferes double chromosome deletion events by relative gene usage.

Usage

```
deletionsByBinom(clip_db, chain = c("IGH", "IGK", "IGL"),
  nonReliable_Vgenes = c())
```

Arguments

`clip_db`	a `data.frame` in Change-O format. See details.
`chain`	the IG chain: IGH,IGK,IGL. Default is IGH.
`nonReliable_Vgenes`	a list of known non reliable gene assignments. A `list` created by `nonReliableVGenes`.

Details

The function accepts a `data.frame` in Change-O format (https://changeo.readthedocs.io/en/version-0.4.1---airr-standards/standard.html) containing the following columns:

- `'SUBJECT'`: The subject name
- `'V_CALL'`: V allele call(s) (in an IMGT format)
- `'D_CALL'`: D allele call(s) (in an IMGT format, only for heavy chains)
- `'J_CALL'`: J allele call(s) (in an IMGT format)

Value

A `data.frame`, in which each row is the double chomosome deletion inference of a gene.

The output containes the following columns:

- `SUBJECT`: the subject name.
- `GENE`: the gene call
- `FRAC`: the relative gene usage of the gene

- CUTOFF: the the cutoff of for the binomial test
- PVAL: the p-value of the binomial test
- DELETION: if a double chromosome deletion event of a gene occured.

Examples

```
# Load example data and germlines
data(samples_db)

# Selecting a single individual
clip_db = samples_db[samples_db$SUBJECT=='I5', ]
# Infering haplotype
del_binom_df = deletionsByBinom(clip_db)
head(del_binom_df)
```

---

deletionsByVpooled *Single chromosomal D or J gene deletions inferred by the V pooled method*

Description

The deletionsByVpooled function inferes single chromosomal deletion for D and J gene .

Usage

```
deletionsByVpooled(clip_db, deletion_col = c("D_CALL"),
  count_thresh = 50, deleted_genes = "", min_minor_fraction = 0.3,
  kThreshDel = 3, nonReliable_Vgenes = c())
```

Arguments

clip_db	a data.frame in Change-O format. See details.
deletion_col	a vector of column names for which single chromosome deletions should be inferred. Default is J_CALL and D_CALL.
count_thresh	integer, the minimun number of sequences mapped to a specific V gene to be included in the V pooled inference.
deleted_genes	double chromosome deletion summary table. A data.frame created by deletionsByBinom.
min_minor_fraction	the minimum minor allele fraction to be used as an anchor gene. Default is 0.3
kThreshDel	the minimum lK (log10 of the Bayes factor) to call a deletion. Defualt is 3.
nonReliable_Vgenes	a list of known non reliable gene assignments. A list created by nonReliableVGenes.

Details

The function accepts a data.frame in Change-O format (https://changeo.readthedocs.io/en/version-0.4.1---airr-standards/standard.html) containing the following columns:

- 'SUBJECT': The subject name
- 'V_CALL': V allele call(s) (in an IMGT format)
- 'D_CALL': D allele call(s) (in an IMGT format, only for heavy chains)
- 'J_CALL': J allele call(s) (in an IMGT format)

Value

A data.frame, in which each row is the single chomosome deletion inference of a gene.

The output containes the following columns:

- SUBJECT: the subject name.
- GENE: the gene call
- DELETION: chromosome deletions inferred. Encoded 1 for deletion and 0 for no deletion.
- K: the Bayesian factor value for the deletion inference.
- COUNTS: the appereance count of the gene on each chromosome, the counts are seperated by a comma.

Examples

```
data(samples_db)

# Infering V pooled deletions
del_db <- deletionsByVpooled(samples_db)
head(del_db)
```

---

hapDendo *Hierarchical clustering of haplotypes graphical output*

Description

The hapDendo function generates a graphical output of an hierarchical clustering based on the Jaccard distance between multiple samples' haplotypes.

Usage

```
hapDendo(hap_table, chain = c("IGH", "IGK", "IGL"),
  gene_sort = c("name", "position"), removeIGH = TRUE,
  mark_low_lk = TRUE, lk_cutoff = 1)
```

Arguments

hap_table	haplotype summary table. See details.
chain	the IG chain: IGH,IGK,IGL. Default is IGH.
gene_sort	if by 'name' the genes in the output are ordered lexicographically, if by 'position' only functional genes are used and are ordered by their chromosomal location. Default is 'position'.
removeIGH	if TRUE, 'IGH'\'IGK'\'IGL' prefix is removed from gene names. Defualt is TRUE.
mark_low_lk	if TRUE, a texture is add for low lK values. Defualt is TRUE.
lk_cutoff	the lK cutoff value to be considerd low for texture layer. Defualt is lK<1.

Details

A `data.frame` created by `createFullHaplotype`.

Value

A multitple samples visualization of the distances between haplotypes.

Examples

```
# Plotting haplotype hierarchical clustering based on the Jaccard distance
hapDendo(samplesHaplotype)
```

---

hapHeatmap *Graphical output of alleles division by chromosome*

Description

The `hapHeatmap` function generates a graphical output of the alleles per gene in multiple samples.

Usage

```
hapHeatmap(hap_table, chain = c("IGH", "IGK", "IGL"),
  gene_sort = "position", removeIGH = TRUE, lk_cutoff = 1,
  mark_low_lk = TRUE)
```

Arguments

hap_table	haplotype summary table. See details.
chain	the IG chain: IGH,IGK,IGL. Default is IGH.
gene_sort	if by 'name' the genes in the output are ordered lexicographically, if by 'position' only functional genes are used and are ordered by their chromosomal location. Default is 'position'.

removeIGH	if TRUE, 'IGH'\'IGK'\'IGL' prefix is removed from gene names.
lk_cutoff	the lK cutoff value to be considerd low for texture layer. Defualt is lK<1.
mark_low_lk	if TRUE, a texture is add for low lK values. Defualt is TRUE.

Details

A data.frame created by createFullHaplotype.

Value

A heat-map visualization of the haplotype inference for multiple samples.

Examples

```
# Plotting haplotpe heatmap
hapHeatmap(samplesHaplotype)
```

HDGERM *Human IGHD germlines*

Description

A character vector of all 37 human IGHD germline gene segment alleles in IMGT Gene-db release 2018-12-4.

Usage

```
HDGERM
```

Format

Values correspond to IMGT nuceltoide sequences.

References

Xochelli *et al*. (2014) Immunoglobulin heavy variable (IGHV) genes and alleles: new entities, new names and implications for research and prognostication in chronic lymphocytic leukaemia. *Immunogenetics*. 67(1):61-6.

HJGERM *Human IGHJ germlines*

Description

A `character` vector of all 13 human IGHJ germline gene segment alleles in IMGT Gene-db release 2018-12-4.

Usage

```
HJGERM
```

Format

Values correspond to IMGT nuceltoide sequences.

References

Xochelli *et al*. (2014) Immunoglobulin heavy variable (IGHV) genes and alleles: new entities, new names and implications for research and prognostication in chronic lymphocytic leukaemia. *Immunogenetics*. 67(1):61-6.

HVGERM *Human IGHV germlines*

Description

A `character` vector of all 342 human IGHV germline gene segment alleles in IMGT Gene-db release 2018-12-4.

Usage

```
HVGERM
```

Format

Values correspond to IMGT-gaped nuceltoide sequences (with nucleotides capitalized and gaps represented by '.').

References

Xochelli *et al*. (2014) Immunoglobulin heavy variable (IGHV) genes and alleles: new entities, new names and implications for research and prognostication in chronic lymphocytic leukaemia. *Immunogenetics*. 67(1):61-6.

nonReliableVGenes *Detect non reliable gene assignment*

Description

nonReliableVGenes Takes a data.frame in Change-O format and detect non reliable IGHV genes. A non reliable gene is when the ratio of the multiple assignments with a gene is below the threshold.

Usage

```
nonReliableVGenes(clip_db, thresh = 0.9, appearance = 0.01)
```

Arguments

clip_db	a data.frame in Change-O format. See details.
thresh	the threshold to consider non reliable gene. Defualt is 0.9
appearance	the minimun fraction of gene appearance to be considered for reliability check. Defualt is 0.01.

Details

The function accepts a data.frame in Change-O format (https://changeo.readthedocs.io/en/version-0.4.1---airr-standards/standard.html) containing the following columns:

- 'SUBJECT': subject names
- 'V_CALL': V allele call(s) (in an IMGT format)

Value

a nested list of non reliable genes for all subject.

Examples

```
# Example IGHV call data frame
clip_db <- data.frame(SUBJECT=rep('S1',6),
V_CALL=c('IGHV1-69*01','IGHV1-69*01','IGHV1-69*01,IGHV1-69*02',
'IGHV4-59*01,IGHV4-61*01','IGHV4-59*01,IGHV4-31*02','IGHV4-59*01'))
# Detect non reliable genes
nonReliableVGenes(clip_db)
```

plotDeletionsByBinom *Graphical output of double chromosome deletions*

Description

The plotDeletionsByBinom function generates a graphical output of the double chromosome deletions in multiple samples.

Usage

```
plotDeletionsByBinom(GENE.usage.df, chain = c("IGH", "IGK", "IGL"),
  genes.low.cer = c("IGHV3-43", "IGHV3-20"), genes.dup = c("IGHD4-11",
  "IGHD5-18"))
```

Arguments

GENE.usage.df	double chromosome deletion summary table. See details.
chain	the IG chain: IGH,IGK,IGL. Default is IGH.
genes.low.cer	a vector of IGH genes known to be with low certantiny in the binomial test. Default is IGHV3-43 and IGHV3-20
genes.dup	a vector of IGH genes known to have a duplicated gene. Default is IGHD4-11 that his duplicate is IGHD4-4 and IGHD5-18 that his duplicate is IGHD5-5

Details

A data.frame created by binom_test_deletion.

Value

A double chromosome deletion visualization.

Examples

```
# Load example data and germlines
data(samples_db)

# Infering haplotype
deletions_db = deletionsByBinom(samples_db);
plotDeletionsByBinom(deletions_db)
```

plotDeletionsByVpooled *Graphical output for single chromosome D or J gene deletions according to V pooled method*

Description

The plotDeletionsByVpooled function generates a graphical output for single chromosome D or J gene deletions (for heavy chain only).

Usage

```
plotDeletionsByVpooled(del.df, K_ranges = c(3, 7))
```

Arguments

del.df	a data.frame created by deletionsByVpooled.
K_ranges	vector of one or two integers for log(K) certainty level thresholds

Details

A data.frame created by deletionsByVpooled.

Value

A single chromosome deletion visualization.

Examples

```
# Load example data and germlines
data(samples_db)
del_db <- deletionsByVpooled(samples_db)
plotDeletionsByVpooled(del_db)
```

plotHaplotype *Graphical output of an inferred haplotype*

Description

The plotHaplotype functions visualizes an inferred haplotype.

Usage

```
plotHaplotype(hap_table, html_output = FALSE, gene_sort = c("name",
  "position"), text_size = 14, removeIGH = TRUE, plotYaxis = TRUE,
  chain = c("IGH", "IGK", "IGL"), dir)
```

Arguments

hap_table	haplotype summary table. See details.
html_output	if TRUE, a html5 interactive graph is outputed. Defualt is FALSE.
gene_sort	if by 'name' the genes in the output are ordered lexicographically, if by 'position' only functional genes are used and are ordered by their chromosomal location. Default is 'position'.
text_size	the size of graph labels. Default is 14 (pts).
removeIGH	if TRUE, 'IGH'\'IGK'\'IGL' prefix is removed from gene names.
plotYaxis	if TRUE, Y axis labels (gene names) are plotted on the middle and right plots. Default is TRUE.
chain	the Ig chain: IGH,IGK,IGL. Default is IGH.
dir	The output folder for saving the haplotype map for multiple individuals.

Details

A `data.frame` in a haplotype format created by `createFullHaplotype` function.

Value

A haplotype map visualization. If more than one subject is visualized, a pdf is created. If html_output is TRUE, a folder named html_output is created with individual graphs.

Examples

```
# Selecting a single individual from the haplotype samples data
haplo_db = samplesHaplotype[samplesHaplotype$SUBJECT=='I5', ]

# plot haplotype
plotHaplotype(haplo_db)
```

rabhit *The RAbHIT package*

Description

The rabhit package provides a robust novel method for determining antibody heavy and light chain haplotypes by adapting a Bayesian framework. The key functions in rabhit, broken down by topic, are described below.

Haplotype and deletions inference

rabhit provides tools to infer haplotypes based on given anchor genes, deletion detection based on relative gene usage, pooling v genes, and a single anchor gene.

- createFullHaplotype: Haplotypes inference and single chromosome deletions based on an anchor gene.
- deletionsByVpooled: Single chromosomal deletion detection by pooling V genes.
- deletionsByBinom: Double chromosomal deletion detection by relative gene usage.
- nonReliableVGenes: Non reliable gene assignment detection.

Haplotype and deletions visualization

Functions for visualization of the inferred haplotypes and deletions

- plotHaplotype: Haplotype inference map.
- deletionHeatmap: Single chromosome deletions heatmap.
- hapHeatmap: Chromosome comparison of multiple samples.
- hapDendo: Hierarchical clustering of multiple haplotypes based on Jaccard distance.
- plotDeletionsByVpooled: V pooled based single chromosome deletions heatmap.
- plotDeletionsByBinom: Double chromosome deletions heatmap.

References

1. Gidoni, M., Snir, O., Peres, A., Polak, P., Lindeman, I., Mikocziova, I., . . . Yaari, G. (2019). Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. Nature Communications, 10(1). doi:10.1038/s41467-019-08489-3

samplesHaplotype *Example haplotype inference results*

Description

A `data.frame` of example haplotype infrence results from createFullHaplotype after double chromosome deletion inference via deletionsByBinom and non reliable V genes detection via nonReliableVGenes. Source data is a colletion of IGH human naive b-cell repertiore data from five individuals (see references). Overall, the data set includes 6 samples. A single individual has two samples (Individual I5), one is short read sequences from BIOMED-2 protocol primers for framework 2 region (The sample is annotated I5_FR2).

Usage

```
samplesHaplotype
```

Format

A `data.frame`, in which each row is the haplotype inference summary of a gene of an individual, from the column selected to prefrom the haplptype infrence on.

References

Gidoni, Moriah, *et al*. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nature Communications*. 10.1 (2019): 628.

See Also

See createFullHaplotype for detailed column descriptions.

samples_db *Example IGH human naive b-cell repertiore*

Description

A `data.frame` of example IGH human naive b-cell repertiore data from five individuals (see references). Overall, the data set includes 6 samples. A single individual has two samples (Individual I5), one is short read sequences from BIOMED-2 protocol primers for framework 2 region (The sample is annotated I5_FR2).

Usage

```
samples_db
```

Format

A `data.frame` in Change-O format (https://changeo.readthedocs.io/en/version-0.4.1---airr-standards/standard.html) containing the following columns:

- `'SUBJECT'`: subject names
- `'V_CALL'`: V allele call(s) (in an IMGT format)
- `'D_CALL'`: D allele call(s) (in an IMGT format, only for heavy chains)
- `'J_CALL'`: J allele call(s) (in an IMGT format)

References

Gidoni, Moriah, *et al*. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nature Communications*. 10.1 (2019): 628.

Index