Package 'SSHAARP'

October 12, 2022

Version 1.1.0

Title Searching Shared HLA Amino Acid Residue Prevalence

Maintainer Livia Tran <livia.tran@ucsf.edu>

Depends R (>= 2.10)

Description Processes amino acid alignments produced by the 'IPD-IMGT/HLA (Immuno Polymorphism-ImMunoGeneTics/Human Leukocyte Antigen) Database' to identify user-defined amino acid residue motifs shared across HLA alleles, and calculates the frequencies of those motifs based on HLA allele frequency data. 'SSHAARP' (Searching Shared HLA Amino Acid Residue Prevalence) uses 'Generic Mapping Tools (GMT)' software and the 'GMT' R package to generate global frequency heat maps that illustrate the distribution of each user-defined map around the globe. 'SSHAARP' analyzes the allele frequency data described by Solberg et al. (2008) <doi:10.1016/j.humimm.2008.05.001>, a global set of 497 population samples from 185 published datasets, representing 66,800 individuals total.

License GPL (>= 3)

Encoding UTF-8

LazyData true

Imports data.table, stringr, gtools, BIGDAWG, gmt, DescTools, dplyr, utils, filesstrings

Suggests knitr, rmarkdown

VignetteBuilder knitr

SystemRequirements GMT (5 or 6), Ghostscript (>=9.6)

RoxygenNote 7.1.1

NeedsCompilation no

Author Livia Tran [aut, cre],
Steven Mack [aut],
Josh Bredeweg [ctb],
Dale Steinhardt [ctb]

Repository CRAN

Date/Publication 2021-09-17 08:30:02 UTC

R topics documented:

AA_atlas 2
BLAASD 3
checkMotif 3
dataSubset 4
findMotif 5
IMGTprotalignments 6
PALM 6
solberg_dataset 8

Index **9**

`AA_atlas` *Exon boundaries for all exons in protein coding genes in the IPD-IMGT/HLA Database release v 3.39.0*

Description

A list object containing exon boundaries for all exons in protein coding genes in the IPD-IMGT/HLA Database release v 3.39.0. Exon boundaries were determined from the nucleotide alignment files, which were downloaded from the ANHIG/IMGTHLA Github respository.

Usage

```
AA_atlas
```

Format

A list containing exon boundaries in a dataframe format for each locus.

Note

For internal use only.

Source

https://github.com/ANHIG/IMGTHLA/tree/Latest/alignments

BLAASD *BLAASD - Build Loci Amino Acid Specific Dataframe*

Description

Extracts alignment sequence information for a given locus from the ANHIG/IMGTHLA GitHub repository to produce a dataframe of individual amino acid data for each amino acid position for all alleles, for a user-defined HLA locus or loci. The first 4 columns are locus, allele, trimmed allele, and allele_name.

Usage

```
BLAASD(loci)
```

Arguments

loci	A vector of un-prefixed HLA locus names

Value

A list object of data frames for each specified locus. Each list element is a data frame of allele names and the corresponding peptide sequence for each amino acid position. An error message is return if the loci input is not a locus for which petpide alignments are available in the ANHIG/IMGTHLA Github Repository.

Examples

```
#BLAASD with one locus as input
BLAASD("C")

#BLAASD with multiple loci as input
BLAASD(c("A", "B", "C"))
```

checkMotif *Syntactic and semantic validation of HLA amino acid motifs*

Description

Checks input motif for errors in format and amino acid positions not present in the locus alignment.

Usage

```
checkMotif(motif)
```

Arguments

motif	An amino acid motif in the following format: Locus*##$~##$~##$, where ## identifies a peptide position, and $ identifies an amino acid residue. Motifs can include any number of amino acids.

Value

A warning message if the input motif is formatted incorrectly, or contains an amino acid position not present in the alignment. Otherwise, a list object with extracted locus information, a correctly formatted motif, and locus specific amino acid dataframe are returned. Note checkMotif() does not check amino acid variants in a specified motif; that is done by findMotif().

Note

For internal SSHAARP use only.

Examples

```
#Example where a motif is formatted correctly
checkMotif("DRB1*26F~28E~30Y")

#Example where format is incorrect
checkMotif("DRB1**26F~28E~30Y")

#Example where an amino acid position does not exist
checkMotif("DRB1**26F~28E~300000Y")
```

---

dataSubset *Solberg dataset manipulation*

Description

Returns a modified version of the Solberg dataset that includes a column of locus*allele names, is sorted by by population name, and is reduced to the specified locus. Cardinal coordinates are converted to their Cartesian equivalents (i.e. 50S is converted to -50).

Usage

```
dataSubset(motif, filename = SSHAARP::solberg_dataset)
```

Arguments

motif	An amino acid motif in the following format: Locus*##$~##$~##$, where ## identifies a peptide position, and $ identifies an amino acid residue. Motifs can include any number of amino acids.
filename	The filename of the local copy of the Solberg dataset - the defaulted filename is the solberg_dataset in the SSHAARP package.

Value

A data frame containing a reformatted version of the Solberg dataset, with rows ordered by population name, Cartesian coordinates in the latit and longit columns, and limited to populations with data for the specified locus. If a motif has formatting errors, a warning message is returned.

Note

For internal SSHAARP use only.

The Solberg dataset is the tab-delimited '1-locus-alleles.dat' text file in the results.zip archive at http://pypop.org/popdata/.

The Solberg dataset is also prepackaged into SSHAARP as 'solberg_dataset'.

`findMotif` *Returns an alignment data frame of alleles that share a specific amino acid motif*

Description

Consumes the alignment data frame produced by BLAASD() and returns an alignment data frame of alleles that share a specific amino acid motif.

Usage

```
findMotif(motif)
```

Arguments

`motif`	An amino acid motif in the following format: Locus*##$~##$~##$, where ## identifies a peptide position, and $ identifies an amino acid residue. Motifs can include any number of amino acids.

Value

An amino acid alignment dataframe of alleles that share the specified motif. If the motif is not found in any alleles, or the motif has formatting errors, a warning message is returned.

Examples

```
#example with actual motif
findMotif("DRB1*26F~28E~30Y")
("DRB1*26F~28E")

#example with non-existent motif
findMotif("DRB1*26F~28E~30Z")

#extracting names of alleles with user-defined motif
findMotif("DRB1*26F~28E~30Y")[,4]
```

IMGTprotalignments — *Protein alignments for all protein coding genes in the IPD-IMGT/HLA Database release v 3.39.0.*

Description

A list object containing protein alignments for all protein coding genes in the IPD-IMGT/HLA Database release. Alignments were downloaded from the ANHIG/IMGTHLA Github respository.

Usage

```
IMGTprotalignments
```

Format

A list containing protein alignments in a dataframe format for each locus.

Source

https://github.com/ANHIG/IMGTHLA/tree/Latest/alignments

PALM — *Population Allele Locating Mapmaker*

Description

Produces a frequency heatmap for a specified amino-acid motif, based on the allele frequency data in the Solberg dataset.

Usage

```
PALM(
  motif,
  filename = SSHAARP::solberg_dataset,
  direct = getwd(),
  color = TRUE,
  filterMigrant = TRUE
)
```

Arguments

motif	An amino acid motif in the following format: Locus*##$~##$~##$, where ## identifies a peptide position, and $ identifies an amino acid residue. Motifs can include any number of amino acids.
filename	The filename of the local copy of the Solberg dataset - the defaulted filename is the solberg_dataset in the SSHAARP package.
direct	The directory into which the map produced is written. The default directory is the user's working directory.
color	A logical parameter that identifies if the heat maps should be made in color (TRUE) or gray scale (FALSE). The default option is TRUE.
filterMigrant	A logical parameter that determines if admixed populations (OTH) and migrant populations (i.e. any complexities with the 'mig') should be excluded from the dataset. The default option is TRUE.

Value

The specified motif and the directory into which the heatmap was written are returned in an invisible character vector. If the user enters a motif that is not found in the Solberg dataset, or that does not exist, a warning message is returned. If an incorrectly formatted motif is entered, or the user does not have the GMT software installed on their operating system, a vector with a warning message is returned. The produced heatmap is written to the user's specified directory (default is user's working directory) as a .jpg file, where the filename is "'motif'.jpg".

Note

The produced frequency heatmap is generated by using the Generic Mapping Tools (GMT) R Package, which is an interface between R and the GMT Map-Making software.

The Solberg dataset is the tab-delimited '1-locus-alleles.dat' text file in the results.zip archive at http://pypop.org/popdata/.

The Solberg dataset is also prepackaged into SSHAARP as 'solberg_dataset'.

While the map legend identifies the highest frequency value, values in this range may not be represented on the map due to frequency averaging over neighboring populations.

References

Solberg et.al. (2008) <doi: 10.1016/j.humimm.2008.05.001>

Examples

```
#example to produce a color frequency heat map where migrant populations are filtered out
PALM("DRB1*26F~28E~30Y",filename = solberg_dataset[85:100,], filterMigrant=TRUE)
#example to produce a greyscale heat map where migrant populations are not filtered out
PALM("DRB1*26F~28E~30Y", filename = solberg_dataset[85:100,], color=FALSE, filterMigrant=FALSE)
```

`solberg_dataset` *Solberg Dataset*

Description

A dataframe of the original Solberg dataset, which is a global dataset of 497 population samples from 185 published datasets, representing 66,800 individuals. For more information on the Solberg dataset, please see the vignette.

Usage

```
solberg_dataset
```

Format

A dataframe with 20163 rows and 13 columns.

Source

results.zip file from http://pypop.org/popdata/

References

Solberg et.al "Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies". Human Immunology (2008) 69, 443–464

Index

∗ **datasets**
AA_atlas, 2
IMGTprotalignments, 6
solberg_dataset, 8

AA_atlas, 2

BLAASD, 3

checkMotif, 3

dataSubset, 4

findMotif, 5

IMGTprotalignments, 6

PALM, 6

solberg_dataset, 8