Package 'ggmanh'

September 5, 2024

Title Visualization Tool for GWAS Result

Version 1.8.0

Description Manhattan plot and QQ Plot are commonly used to visualize the end result of Genome Wide Association Study.
The ``ggmanh" package aims to keep the generation of these plots simple while maintaining customizability.
Main functions include manhattan_plot, qqunif, and thinPoints.

biocViews Visualization, GenomeWideAssociation, Genetics

License MIT + file LICENSE

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.2

Imports gdsfmt, ggrepel, grDevices, RColorBrewer, rlang, scales, SeqArray (>= 1.32.0), stats

Depends methods, ggplot2

Suggests BiocStyle, rmarkdown, knitr, testthat (>= 3.0.0), markdown, GenomicRanges

Config/testthat/edition 3

VignetteBuilder knitr

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

git_branch RELEASE_3_19

git_last_commit 5f3d316

git_last_commit_date 2024-04-30

Repository Bioconductor 3.19

Date/Publication 2024-09-04

Author John Lee [aut, cre],
John Lee [aut] (AbbVie),
Xiuwen Zheng [ctb, dtc]

Maintainer John Lee <swannyy.stat@gmail.com>

Contents

`default_gds_path` *Path to Default GDS File*

Description

Find path to the default gds file.

Usage

```
default_gds_path()
```

Value

A character vector.

Examples

```
default_gds_path()
```

`gds_annotate` *Annotation with GDS File*

Description

Retrieve variant annotation stored in a GDS file with chromosome location or rs.id.

Usage

```
gds_annotate(
  x,
  gdsfile = NULL,
  annot.method = "position",
  chr = NULL,
  pos = NULL,
  ref = NULL,
  alt = NULL,
  rs.id = NULL,
  concat_char = "/",
  verbose = TRUE,
  annotation_names = c("annotation/info/symbol", "annotation/info/consequence",
    "annotation/info/LoF")
)
```

Arguments

x	a data.frame object to be annotated.
gdsfile	a character for GDS filename. If NULL, the default GDS file included with the package is used.
annot.method	a method for searching variants. "position" requires chr, pos, ref, and alt. "rs.id" requires rs.id.
chr, pos, ref, alt, rs.id	column names of x that contain chromosome, position, reference allele, alternate allele, and rs.id, respectively.
concat_char	a character used to separate multiple annotations returned from the gds file.
verbose	output messages.
annotation_names	a character vector of nodes of the gdsfile that are to be extracted.

Value

A character vector the length of nrow(x) if concat_char is a character. A data frame with nrow(x) rows and length(annotation_names) if concat_char is null.

Examples

```
vardata <- data.frame(
  chr = c(11,20,14),
  pos = c(12261002, 10033792, 23875025),
  ref = c("G", "G", "CG"),
  alt = c("A", "A", "C")
)

annotations <- gds_annotate(
  x = vardata, annot.method = "position",
  chr = "chr", pos = "pos", ref = "ref", alt = "alt"
```

```
)

print(annotations)
```

---

ggmanh *ggmanh: A package for visualization of GWAS results.*

Description

ggmanh provides flexible tools for visualizing GWAS result for downstream analysis.

Details

Manhattan plot is commonly used to display significant Single Nucleotide Polymorphisms (SNPs) in Genome Wide Association Study (GWAS) This package comes with features useful for manhattan plot creation, including annotation with `ggrepel`, truncating data for faster plot generation, and manual rescaling of the y-axis. The manhattan plot is generated in two steps: data preprocessing and plotting. This allows the user to iteratively customize the plot without having the process the GWAS summary data over and over again. Currently, `data.frame` and `GRanges` from `GenomicRanges` are supported.

A vignette detailing the usage of the package is accessible by `vignette("ggmanh")`

---

ggmanh_annotation_gds *gnomAD Variant Annotation in SeqArray Format*

Description

`ggmanh` provides a GDS file whose path is accessible by `default_gds_path`. The original annotation file is from the gnomAD browser v2.1.1 release, available in this link: https://gnomad.broadinstitute.org/downloads. This gds file contains variants in the exome with the global minor allele frequency ≥ 0.0002, and has been manually curated to fit the file size requirement for R Bioconductor packages.

Format

A GDS file with 1015430 variants with chromosome, position, allele, gene symbol, Ensembl VEP Consequence, and predicted LoF.

manhattan_data_preprocess *Preprocess GWAS Result*

Description

Preprocesses a result from Genome Wide Association Study before making a manhattan plot. It accepts a `data.frame`, which at bare minimum should contain a chromosome, position, and p-value. Additional options, such as chromosome color, label colum names, and colors for specific variants, are provided here.

Usage

```
manhattan_data_preprocess(x, ...)

## Default S3 method:
manhattan_data_preprocess(x, ...)

## S3 method for class 'data.frame'
manhattan_data_preprocess(
  x,
  chromosome = NULL,
  signif = c(5e-08, 1e-05),
  pval.colname = "pval",
  chr.colname = "chr",
  pos.colname = "pos",
  highlight.colname = NULL,
  chr.order = NULL,
  signif.col = NULL,
  chr.col = NULL,
  highlight.col = NULL,
  preserve.position = FALSE,
  thin = NULL,
  thin.n = 1000
)

## S4 method for signature 'GRanges'
manhattan_data_preprocess(
  x,
  chromosome = NULL,
  signif = c(5e-08, 1e-05),
  pval.colname = "pval",
  highlight.colname = NULL,
  chr.order = NULL,
  signif.col = NULL,
  chr.col = NULL,
  highlight.col = NULL,
```

```
  preserve.position = FALSE,
  thin = NULL,
  thin.n = 100
)
```

Arguments

x	a data frame or any other extension of data frame (e.g. a tibble). At bare minimum, it should contain chromosome, position, and p-value.
...	Additional arguments for manhattan_data_preprocess.
chromosome	a character. This is supplied if a manhattan plot of a single chromosome is desired. If NULL, then all the chromosomes in the data will be plotted.
signif	a numeric vector. Significant p-value thresholds to be drawn for manhattan plot. At least one value should be provided. Default value is c(5e-08, 1e-5)
pval.colname	a character. Column name of x containing p.value.
chr.colname	a character. Column name of x containing chromosome number.
pos.colname	a character. Column name of x containing position.
highlight.colname	a character. If you desire to color certain points (e.g. significant variants) rather than color by chromosome, you can specify the category in this column, and provide the color mapping in highlight.col. Ignored if NULL.
chr.order	a character vector. Order of chromosomes presented in manhattan plot.
signif.col	a character vector of equal length as signif. It contains colors for the lines drawn at signif. If NULL, the smallest value is colored black while others are grey.
chr.col	a character vector of equal length as chr.order. It contains colors for the chromosomes. Name of the vector should match chr.order. If NULL, default colors are applied using RColorBrewer.
highlight.col	a character vector. It contains color mapping for the values from highlight.colname.
preserve.position	a logical. If TRUE, the width of each chromosome reflect the number of variants and the position of each variant is correctly scaled? If FALSE, the width of each chromosome is equal and the variants are equally spaced.
thin	a logical. If TRUE, thinPoints will be applied. Defaults to TRUE if chromosome is NULL. Defaults to FALSE if chromosome is supplied.
thin.n	an integer. Number of max points per horizontal partitions of the plot. Defaults to 1000.

Details

manhattan_data_preprocess gathers information needed to plot a manhattan plot and organizes the information as MPdata S3 object.

New positions for each points are calculated, and stored in the data.frame as "new_pos". By default, all chromosomes will have the same width, with each point being equally spaced. This behavior

is changed when `preserve.position = TRUE`. The width of each chromosome will scale to the number of points and the points will reflect the original positions.

`chr.col` and `highlight.col`, maps the data values to colors. If they are an unnamed vector, then the function will try its best to match the values of `chr.colname` or `highlight.colname` to the colors. If they are a named vector, then they are expected to map all values to a color. If `highlight.colname` is supplied, then `chr.col` is ignored.

While feeding a `data.frame` directly into `manhattan_plot` does preprocessing & plotting in one step. If you plan on making multiple plots with different graphic options, you have the choice to preprocess separately and then generate plots.

Value

a MPdata object. This object contains all the necessary info for constructing a manhattan plot.

Examples

```
gwasdat <- data.frame(
  "chromosome" = rep(1:5, each = 30),
  "position" = c(replicate(5, sample(1:300, 30))),
  "pvalue" = rbeta(150, 1, 1)^5
)

  manhattan_data_preprocess(
  gwasdat, pval.colname = "pvalue", chr.colname = "chromosome", pos.colname = "position",
  chr.order = as.character(1:5)
)
```

`manhattan_plot` *Manhattan Plotting*

Description

A generic function for manhattan plot.

Usage

```
manhattan_plot(x, ...)

manhattan_plot.default(x, ...)

## S3 method for class 'data.frame'
manhattan_plot(
  x,
  chromosome = NULL,
  outfn = NULL,
  signif = c(5e-08, 1e-05),
  pval.colname = "pval",
```

```
  chr.colname = "chr",
  pos.colname = "pos",
  label.colname = NULL,
  highlight.colname = NULL,
  chr.order = NULL,
  signif.col = NULL,
  chr.col = NULL,
  highlight.col = NULL,
  rescale = TRUE,
  rescale.ratio.threshold = 5,
  signif.rel.pos = 0.4,
  color.by.highlight = FALSE,
  preserve.position = FALSE,
  thin = NULL,
  thin.n = 1000,
  plot.title = ggplot2::waiver(),
  plot.subtitle = ggplot2::waiver(),
  plot.width = 10,
  plot.height = 5,
  point.size = 0.75,
  label.font.size = 2,
  max.overlaps = 20,
  x.label = "Chromosome",
  y.label = expression(-log[10](p)),
  ...
)

## S3 method for class 'MPdata'
manhattan_plot(
  x,
  chromosome = NULL,
  outfn = NULL,
  rescale = TRUE,
  rescale.ratio.threshold = 5,
  signif.rel.pos = 0.4,
  color.by.highlight = FALSE,
  label.colname = NULL,
  x.label = "Chromosome",
  y.label = expression(-log[10](p)),
  point.size = 0.75,
  label.font.size = 2,
  max.overlaps = 20,
  plot.title = ggplot2::waiver(),
  plot.subtitle = ggplot2::waiver(),
  plot.width = 10,
  plot.height = 5,
  ...
)
```

```
## S4 method for signature 'GRanges'
manhattan_plot(
  x,
  chromosome = NULL,
  outfn = NULL,
  signif = c(5e-08, 1e-05),
  pval.colname = "pval",
  label.colname = NULL,
  highlight.colname = NULL,
  chr.order = NULL,
  signif.col = NULL,
  chr.col = NULL,
  highlight.col = NULL,
  rescale = TRUE,
  rescale.ratio.threshold = 5,
  signif.rel.pos = 0.4,
  color.by.highlight = FALSE,
  preserve.position = FALSE,
  thin = NULL,
  thin.n = 1000,
  plot.title = ggplot2::waiver(),
  plot.subtitle = ggplot2::waiver(),
  plot.width = 10,
  plot.height = 5,
  point.size = 0.75,
  label.font.size = 2,
  max.overlaps = 20,
  x.label = "Chromosome",
  y.label = expression(-log[10](p)),
  ...
)
```

Arguments

x	a data.frame, an extension of data.frame object (e.g. tibble), or an MPdata object.
...	additional arguments to be passed onto geom_label_repel
chromosome	a character. This is supplied if a manhattan plot of a single chromosome is desired. If NULL, then all the chromosomes in the data will be plotted.
outfn	a character. File name to save the Manhattan Plot. If outfn is supplied (i.e. !is.null(outfn)), then the plot is not drawn in the graphics window.
signif	a numeric vector. Significant p-value thresholds to be drawn for manhattan plot. At least one value should be provided. Default value is c(5e-08, 1e-5)
pval.colname	a character. Column name of x containing p.value.
chr.colname	a character. Column name of x containing chromosome number.
pos.colname	a character. Column name of x containing position.

label.colname a character. Name of the column in MPdata$data to be used for labelling.

highlight.colname
a character. If you desire to color certain points (e.g. significant variants) rather than color by chromosome, you can specify the category in this column, and provide the color mapping in highlight.col. Ignored if NULL.

chr.order a character vector. Order of chromosomes presented in manhattan plot.

signif.col a character vector of equal length as signif. It contains colors for the lines drawn at signif. If NULL, the smallest value is colored black while others are grey.

chr.col a character vector of equal length as chr.order. It contains colors for the chromosomes. Name of the vector should match chr.order. If NULL, default colors are applied using RColorBrewer.

highlight.col a character vector. It contains color mapping for the values from highlight.colname.

rescale a logical. If TRUE, the plot will rescale itself depending on the data. More on this in details.

rescale.ratio.threshold
a numeric. Threshold of that triggers the rescale.

signif.rel.pos a numeric between 0.1 and 0.9. If the plot is rescaled, where should the significance threshold be positioned?

color.by.highlight
a logical. Should the points be colored based on a highlight column?

preserve.position
a logical. If TRUE, the width of each chromosome reflect the number of variants and the position of each variant is correctly scaled? If FALSE, the width of each chromosome is equal and the variants are equally spaced.

thin a logical. If TRUE, thinPoints will be applied. Defaults to TRUE if chromosome is NULL. Defaults to FALSE if chromosome is supplied.

thin.n an integer. Number of max points per horizontal partitions of the plot. Defaults to 1000.

plot.title a character. Plot title

plot.subtitle a character. Plot subtitle

plot.width a numeric. Plot width in inches.

plot.height a numeric. Plot height in inches.

point.size a numeric. Size of the points.

label.font.size
a numeric. Size of the labels.

max.overlaps an integer. Exclude text labels that overlaps too many things.

x.label a character. x-axis label

y.label a character. y-axis label

Details

This generic function accepts a result of a GWAS in the form of `data.frame` or a `MPdata` object produced by `manhattan_data_preprocess`. The function will throw an error if another type of object is passed.

Having `rescale = TRUE` is useful when there are points with very high -log10(p.value). In this case, the function attempts to split the plot into two different scales, with the split happening near the strictest significance threshold. More precisely, the plot is rescaled when

$$-log10(pvalue)/(strictestsignificancethreshold) \geq rescale.ratio.threshold$$

If you wish to add annotation to the points, provide the name of the column to `label.colname`. The labels are added with `ggrepel`.

Be careful though: if the annotation column contains a large number of variants, then the plotting could take a long time, and the labels will clutter up the plot. For those points with no annotation, you have the choice to set them as `NA` or `""`.

Value

gg object if `is.null(outfn)`, `NULL` if `!is.null(outf)`

Examples

```
gwasdat <- data.frame(
  "chromosome" = rep(1:5, each = 30),
  "position" = c(replicate(5, sample(1:300, 30))),
  "pvalue" = rbeta(150, 1, 1)^5
)

manhattan_plot(
  gwasdat, pval.colname = "pvalue", chr.colname = "chromosome", pos.colname = "position",
  chr.order = as.character(1:5)
)

mpdata <- manhattan_data_preprocess(
  gwasdat, pval.colname = "pvalue", chr.colname = "chromosome", pos.colname = "position",
  chr.order = as.character(1:5)
)

manhattan_plot(mpdata)
```

---

qqunif *Plot Quantile-Quantile Plot of p-values against uniform distribution.*

Description

Plot Quantile-Quantile Plot of p-values against uniform distribution.

Usage

```
qqunif(
  x,
  outfn = NULL,
  conf.int = 0.95,
  plot.width = 5,
  plot.height = 5,
  thin = TRUE,
  thin.n = 500,
  zero.pval = "replace"
)
```

Arguments

x	a numeric vector of p-values. All values should be between 0 and 1.
outfn	a character. File name to save the QQ Plot. If outfn is supplied (i.e. !is.null(outfn)), then the plot is not drawn in the graphics window.
conf.int	a numeric between 0 and 1. Confidence band to draw around reference line. Set to NA to leave it out.
plot.width	a numeric. Plot width in inches.
plot.height	a numeric. Plot height in inches.
thin	a logical. Reduce number of data points when they are cluttered?
thin.n	an integer. Number of max points per horizontal partitions of the plot. Defaults to 500.
zero.pval	a character. Determine how to treat 0 pvals. "replace" will replace the p-value of zero with the non-zero minimum. "remove" will remove the p-value of zero.

Value

a ggplot object

Examples

```
x <- rbeta(1000, 1, 1)
qqunif(x)
```

---

thinPoints *Thin Data Points*

Description

Reduce the number of cluttered data points.

Usage

```
thinPoints(dat, value, n = 3000, nbins = 200, groupBy = NULL)
```

Arguments

dat	a data frame
value	column name of dat to be used for partitioning (see details)
n	number of points to sample for each partition
nbins	number of partitions
groupBy	column name of dat to group by before partitioning (e.g. chromosome)

Details

The result of Genome Wide Association Study can be very large, with the majority of points being being clustered below significance threshold. This unnecessarily increases the time to plot while making almost no difference. This function reduces the number of points by partitioning the points by a numberic column value into nbins and sampling n points.

Value

a data.frame

Examples

```
dat <- data.frame(
   A1 = c(1:20, 20, 20),
   A2 = c(rep(1, 12), rep(1,5), rep(20, 3), 20, 20) ,
   B = rep(c("a", "b", "c", "d"), times = c(5, 7, 8, 2))
)
# partition "A1" into 2 bins and then sample 6 data points
thinPoints(dat, value = "A1", n = 6, nbins = 2)
# partition "A2" into 2 bins and then sample 6 data points
thinPoints(dat, value = "A2", n = 6, nbins = 2)
# group by "B", partition "A2" into 2 bins and then sample 3 data points
thinPoints(dat, value = "A2", n = 3, nbins = 2, groupBy = "B")
```

Index