Package 'garfield'

October 17, 2017

Type Package

Title GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction

Version 1.4.0

Date 2015-12-14

Author Sandro Morganella <sm22@sanger.ac.uk>

Maintainer Valentina Iotchkova <vi1@sanger.ac.uk>

Description GARFIELD is a non-parametric functional enrichment analysis approach described in the paper GARFIELD: GWAS analysis of regulatory or functional information enrichment with LD correction. Briefly, it is a method that leverages GWAS findings with regulatory or functional annotations (primarily from ENCODE and Roadmap epigenomics data) to find features relevant to a phenotype of interest. It performs greedy pruning of GWAS SNPs (LD r2 > 0.1) and then annotates them based on functional information overlap. Next, it quantifies Fold Enrichment (FE) at various GWAS significance cutoffs and assesses them by permutation testing, while matching for minor allele frequency, distance to nearest transcription start site and number of LD proxies (r2 > 0.8).

biocViews Software, StatisticalMethod, Annotation, FunctionalPrediction, GenomeAnnotation

License GPL-3

NeedsCompilation yes

VignetteBuilder knitr

Suggests knitr

R topics documented:

`garfield-package` *GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction*

Description

GARFIELD leverages GWAS findings with regulatory or functional annotations to find features relevant to a phenotype of interest. It performs greedy pruning of GWAS SNPs (LD r2>0.1) and then annotates them based on functional information overlap. Next, it quantifies Fold Enrichment (FE) at various GWAS signficance cut-offs and assesses them by permutation testing, while matching for minor allele frequency, distance to nearest transcription start site and number of LD proxies (r2>0.8). Finally, it includes visualization tools.

Details

Package:	garfield
Type:	Package
Version:	1.0
Date:	2015-12-14
License:	GPL-3

See garfield.run for example analysis usage and garfield.plot for plotting examples.

Author(s)

Sandro Morganella <email: sm22@sanger.ac.uk>

Maintainer: Valentina Iotchkova <email: vi1@sanger.ac.uk>

References

Valentina Iotchkova, Graham Ritchie, Matthias Geihs, Sandro Morganella, Josine Min, Klaudia Walter, Nicholas Timpson, UK10K Consortium, Ian Dunham, Ewan Birney and Nicole Soranzo. GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. In preparation

See Also

`garfield.run`, `garfield.plot`

`garfield.plot` *Garfield plotting function*

Description

`garfield.plot` is used for visualization of the enrichment analysis results obtained by the `garfield.run` permutation step. Internally, it uses `garfield.plot.fnx` which has been adapted from the 'radial.plot' function from the 'plotrix' package.

Usage

```
garfield.plot(input_file, num_perm = 100000, output_prefix = "plot",
    plot_title = "", filter = 10, tr = Inf)
```

Arguments

input_file	Input file name as produced by garfield.run permutation step.
num_perm	Number of permutations used in the garfield.run permutation step.
output_prefix	Figure file prefix. This would create the following files output_prefix.Chromatin_States.pdf, output_prefix.Footprints.pdf, output_prefix.Histone_Modifications.pdf, output_prefix.Peaks.pdf, output_prefix.FAIRE.pdf, output_prefix.Genic.pdf, output_prefix.Hotspots.pdf and output_prefix.TFBS.pdf
plot_title	Optional figure title
filter	Optional filter for the minimum number of variants at a given threshold. Minimum of 1 should be used, but advisable to set to a larger value (e.g. 10).
tr	Threshold for denoting significance of an observed enrichment on a -log10 scale. A value of Inf denotes using default threshold of -log10(0.05/498).

Details

This function is used for visualization of the enrichment analysis results and produces pdf figures for each class of annotations. Each figure shows the FE values (radial values) at different GWAS thresholds (bottom legend) for each annotation (outer circle and right legend). It further shows the significance at the top 4 GWAS thresholds (if present) as dots on the outer circle, with the most stringent threshold being shown at the inner most side.

Value

No value is produced, instead output files are generated. See Details and 'output_prefix' for more information.

Author(s)

Sandro Morganella <email: sm22@sanger.ac.uk>

Maintainer: Valentina Iotchkova <email: vi1@sanger.ac.uk>

References

Valentina Iotchkova, Graham Ritchie, Matthias Geihs, Sandro Morganella, Josine Min, Klaudia Walter, Nicholas Timpson, UK10K Consortium, Ian Dunham, Ewan Birney and Nicole Soranzo. GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. In preparation

See Also

garfield.run, garfield, garfield.plot.fnx

Examples

```
garfield.run("tmp", data.dir=system.file("extdata",package = "garfield"),
    trait="trait",run.option = "prep", chrs = c(22),
    exclude = c(895, 975, 976, 977, 978, 979, 98))

garfield.run("tmp", data.dir=system.file("extdata",package = "garfield"),
    trait="", run.option = "perm", nperm = 1000,
    thresh = c(0.001, 1e-04, 1e-05), pt_thresh = c(1e-04, 1e-05),
    maf.bins = 2, tags.bins = 3, tss.bins = 3, prep.file = "tmp.prep",
    optim_mode = TRUE, minit = 100, thresh_perm = 0.05)

garfield.plot("tmp.perm", num_perm = 1000, output_prefix = "tmp",
    plot_title = "Sample run", filter = 1, tr = -log10(0.05))
#system("ls -lh tmp.*.pdf")
```

---

garfield.plot.fnx *Internal radial plotting function for the garfield.plot function*

Description

This function has been adapted from the 'radial.plot' function from the 'plotrix' package.

Usage

```
garfield.plot.fnx(lengths, radial.pos = NULL, labels = NA, breaks = NA,
    label.pos = NULL, radlab = FALSE, start = 0, clockwise = FALSE,
    rp.type = "r", label.prop = 1.05, main = "", xlab = "", ylab = "",
    line.col = par("fg"), lty = par("lty"), lwd = par("lwd"),
    mar = c(2, 2, 3, 2), show.grid = TRUE, show.grid.labels = 4,
    show.radial.grid = TRUE, grid.col = "grey", grid.bg = "transparent",
    grid.left = FALSE, grid.unit = NULL, point.symbols = 1,
    point.col = par("fg"), show.centroid = FALSE, radial.lim = NULL,
    radial.labels = NULL, poly.col = NA, add = FALSE, ann.col = 1,
    ann.pch = 15, ann.col.mx = 1, compact = TRUE, ...)
```

Arguments

lengths	A numeric data vector or matrix. If 'lengths' is a matrix, the rows will be considered separate data vectors.
radial.pos	A numeric vector or matrix of positions in radians. These are interpreted as beginning at the right (0 radians) and moving counterclockwise. If 'radial.pos' is a matrix, the rows must correspond to rows of 'lengths'.
labels	Character strings to be placed at the outer ends of the lines. If set to NA, will suppress printing of labels, but if missing, the radial positions will be used.
breaks	A vector of (potentially different) labels to 'labels' according to which to draw radial lines.
label.pos	The positions of the labels around the plot in radians.
radlab	Whether to rotate the outer labels to a radial orientation.

start	Where to place the starting (zero) point. Defaults to the 3 o'clock position.
clockwise	Whether to interpret positive positions as clockwise from the starting point. The default is counterclockwise.
rp.type	Whether to draw (r)adial lines, a (p)olygon, (s)ymbols or some combination of these. If 'lengths' is a matrix and rp.type is a vector, each row of 'lengths' can be displayed differently.
label.prop	The label position radius as a proportion of the maximum line length.
main	The title for the plot.
xlab,ylab	Normally x and y axis labels are suppressed.
line.col	The color of the radial lines or polygons drawn.
lty	The line type(s) to be used for polygons or radial lines.
lwd	The line width(s) to be used for polygons or radial lines.
mar	Margins for the plot. Allows the user to leave space for legends, long labels, etc.
show.grid	Logical - whether to draw a circular grid.
show.grid.labels	Whether and where to display labels for the grid - see Details.
show.radial.grid	Whether to draw radial lines to the plot labels.
grid.col	Color of the circular grid.
grid.bg	Fill color of above.
grid.left	Whether to place the radial grid labels on the left side.
grid.unit	Optional unit description for the grid.
point.symbols	The symbols for plotting (as in pch).
point.col	Colors for the symbols.
show.centroid	Whether to display a centroid.
radial.lim	The range of the grid circle. Defaults to 'pretty(range(lengths))', but if more than two values are passed, the exact values will be displayed.
radial.labels	Optional labels for the radial grid. The default is the values of radial.lim.
poly.col	Fill color if polygons are drawn. Use NA for no fill.
add	Whether to add one or more series to an existing plot.
ann.col	A vector of colours for symbols to be drawn on the outer circle.
ann.pch	A pch value for symbols to be drawn on the outer circle.
ann.col.mx	A matrix of colours for dots to be drawn on the outer circle, inside the symbols with 'ann.col' colours.
compact	A flag to specify if labels should be presented in a compact form or not.
...	Additional arguments are passed to 'plot'.

Details

This function is only used internally as part of the garfield.plot function.

Value

No value, but a plot is produced. See Details for further information.

Author(s)

Sandro Morganella <email: sm22@sanger.ac.uk>

Maintainer: Valentina Iotchkova <email: vi1@sanger.ac.uk>

References

Valentina Iotchkova, Graham Ritchie, Matthias Geihs, Sandro Morganella, Josine Min, Klaudia Walter, Nicholas Timpson, UK10K Consortium, Ian Dunham, Ewan Birney and Nicole Soranzo. GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. In preparation

See Also

garfield.plot, garfield.run, garfield

Examples

```
DATA = rbind(rnorm(10,5,0.5),rnorm(10,3.5,0.5),rnorm(10,2,0.5))
garfield.plot.fnx(DATA,ann.col.mx=DATA!=0, ann.col=rep(1:2,each=5),
    ann.pch=15, rp.type="p",line.col=1:3,show.grid=TRUE, show.radial.grid=TRUE,
    labels=paste("label ",c(1:10)," ",sep=""),breaks=(1:10), radlab=TRUE,
    poly.col=1:3)
```

---

garfield.run *GARFIELD enrichment analysis function*

Description

garfield.run is used to perform greedy pruning of variants from a genome-wide association study, calculate fold enrichment and test its significance at a given genome-wide significance threshold.

Usage

```
garfield.run(out.file, data.dir, trait, run.option = "complete",
    chrs = c(1:22, "X"), exclude = c(895, 975, 976, 977, 978, 979, 98),
    nperm = 100000, thresh = c(0.1, 0.01, 0.001, 1e-04, 1e-05, 1e-06,
    1e-07, 1e-08), pt_thresh = c(1e-05, 1e-06, 1e-07, 1e-08),
    maf.bins = 5, tags.bins = 5, tss.bins = 5, prep.file = "",
    optim_mode = 1, minit = 100, thresh_perm = 1e-04)
```

Arguments

out.file	Prefix for output file. Full garfield analysis creates out.file.prep and out.file.perm files from prunning and annotation step and permutation for significance testing step, respectively. These steps can additionally be run separately - for more information see run.option flag.
data.dir	Path to annotation and p-value files. The directory must contain "annotation","maftssd","pval" and "tags" subdirectories with per chromosome files of input data. See details section for further information.

`trait`	GWAS phenotype name. This must match a folder name in the data.dir folder. See details for more information.
`run.option`	an object specifying which part of the analysis to run. Valid options are `complete`, `prep` and `perm`, where `prep` denotes the preparation step (pruning and annotation of variants), `perm` denotes the permutation step (calculating fold enrichment and its significance) and `complete` executes both the preparation and permutation steps.
`chrs`	A vector of the chromosomes for which to run the enrichment analysis. `chrs` can have all or subsets of values from `c(1:22,'X')`.
`exclude`	A numeric vector of indeces of annotations for which LD tags should not be used for annotation of pruned variants. Value of -1 denotes using LD for annotation for all features.
`nperm`	A numeric value of the number of permutations to be performed.
`thresh`	A numeric vector of genome-wide significance thresholds to be used for fold enrichment calculation.
`pt_thresh`	A numeric vector of genome-wide significance thresholds to be used for calculating the significance of the observed fold enrichment. All values must be contained in the `thresh` vector.
`maf.bins`	A numeric value denoting the number of bins for the minor allele frequency matching during permutation testing. Must be greater or equal to 1.
`tags.bins`	A numeric value denoting the number of bins for the number of LD tags (r2>=0.8) matching during permutation testing. Must be greater or equal to 1.
`tss.bins`	A numeric value denoting the number of bins for the distance to nearest transcription start site matching during permutation testing. Must be greater or equal to 1.
`prep.file`	File from pruning stage of algorithm. Only required if using the 'run.option'=perm flag.
`optim_mode`	A binary flag denoting whether to run fast version of method (1) or general version (0), where the fast version checks if significance of a given enrichment would still be possible to be obtained after 'minit' number of iterations and terminated permutations if not.
`minit`	An integer value for the minimum number of permutations to be performed before checking if `thresh_perm` condition can still be met. Only used if 'optim_mode'=1.
`thresh_perm`	After 'minit' number of permutations, at each iteration check if EmpPval can still reach 'thresh_perm' value. If not terminate permutations and output obtained results at that stage. Only used if 'optim_mode'=1.

Details

Output files: out.file.prep contains the genomic positions of pruned variants, p-values for association with the trait of interest, number of LD tags (r2>0.8), MAF, distance to the nearest TSS and binary representation of annotation information (with LD-tagging r2>0.8). out.file.perm contains enrichment analysis results for each annotation, where PThresh is the GWAS p-value threshold used for analysis, FE denotes the fold enrichment statistic (equals -1 if no sufficient data was available for the FE calculation), EmpPval shows the empirical p-value of enrichment (equals -1 if FE is calculated but significance of enrichment analysis is not run at that threshold), NAnnotThresh - the number of variants at the threshold which are annotated with the given feature, NAnnot - the total number of annotated variants, NThresh - the total number of variants at that threshold and N

- the total number of pruned variants. The remaining columns show additional information on the annotations used for analysis.

Data directory: data.dir, should point to a location containing "annotation","maftssd","pval" and "tags" subdirectories, where (i) the "pval" folder contains subfolders with trait names, which in turn contain per chromosome space separated files with genomic position in the first column and p-value in the second. They should be named chr1, chr2, etc, and be numerically sorted with respect to genomic position; (ii) "annotation" should contain per chromosome space separated files with position in the first column and annotations in stacked binary format in the second. The files should be named chr1, chr2, etc and be sorted numerically according to position. Additionally the directory should contain a link_file.txt file that links the annotations to relevant information about them; (iii) "maftssd" should contain per chromosome space separated files with position in the first column, minor allele frequency in the second and distance to the nearest TSS in the third. The files should be named chr1, chr2, etc and be sorted numerically according to position; (iv) "tags" should contain two subfolders named r01 and r08, which in turn contain per chromosome space separated files with variant position in the first column and comma separated positions of all variants with r2>0.1 or 0.8, respectively with the variant in the first column. The files should again be named chr1, chr2, etc and be sorted numerically according to position. For further information on the data.dir structure, please see http://www.ebi.ac.uk/birney-srv/GARFIELD/documentation/GARFIELD.pdf

Pre-computed LD (European samples - UK10K sequence data), MAF, TSS distance, p-value files for two example traits (Crohn's Disease from the IBD Consortium and Height from the GIANT consortium) and annotation files for 1005 GENCODE, ENCODE and Roadmap Epigenomics annotations can be downloaded from http://www.ebi.ac.uk/birney-srv/GARFIELD/package/garfield-data.tar.gz. Note the data is 5.9Gb in compressed format and needs to be uncompressed prior to analysis (83Gb). Variant genomic position (build 37) is used as an identifier in all data files.

Value

No value is produced, instead output files are generated. See Details and 'out.file' for more information.

Author(s)

Sandro Morganella <email: sm22@sanger.ac.uk>

Maintainer: Valentina Iotchkova <email: vi1@sanger.ac.uk>

References

Valentina Iotchkova, Graham Ritchie, Matthias Geihs, Sandro Morganella, Josine Min, Klaudia Walter, Nicholas Timpson, UK10K Consortium, Ian Dunham, Ewan Birney and Nicole Soranzo. GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. In preparation

See Also

garfield.plot, garfield

Examples

```
garfield.run("tmp", data.dir=system.file("extdata",package = "garfield"),
    trait="trait",run.option = "prep", chrs = c(22),
    exclude = c(895, 975, 976, 977, 978, 979, 98))
```

```
garfield.run("tmp", data.dir=system.file("extdata",package = "garfield"),
    run.option = "perm", nperm = 1000, thresh = c(0.001, 1e-04, 1e-05),
    pt_thresh = c(1e-04, 1e-05), maf.bins = 2, tags.bins = 3, tss.bins = 3,
    prep.file = "tmp.prep", optim_mode = TRUE, minit = 100, thresh_perm = 0.05)

if (file.exists("tmp.perm")){
    perm = read.table("tmp.perm", header=TRUE)
    head(perm)
} else { print("Error: tmp.perm does not exist!") }

###### To get the sample data for enrichment analysis in European samples
###### execute the following - note this can take a long time to run and
###### needs a substantial disk space (see Details)
#
# download data and decompress
# system("wget http://www.ebi.ac.uk/birney-srv/GARFIELD/package/
# garfield-data.tar.gz")
# system("tar -zxvf garfield-data.tar.gz")
#
# if downloaded in current working directory use the following to execute
# garfield, otherwise please change data.dir location
# garfield.run("cd-meta.output", data.dir="garfield-data", trait="cd-meta",
# run.option = "prep", chrs = c(1:22), exclude = c(895, 975, 976, 977, 978,
# 979, 980))
#
# garfield.run("cd-meta.output", data.dir="garfield-data", run.option = "perm",
# nperm = 100000, thresh = c(0.1,0.01,0.001, 1e-04, 1e-05, 1e-06, 1e-07, 1e-08),
# pt_thresh = c(1e-05, 1e-06, 1e-07, 1e-08), maf.bins = 5, tags.bins = 5,
# tss.bins = 5, prep.file = "cd-meta.output.prep", optim_mode = TRUE,
# minit = 100, thresh_perm = 0.0001)
#
# garfield.plot("cd-meta.output.perm", num_perm = 100000,
# output_prefix = "cd-meta.output", plot_title = "Crohn's Disease",
# filter = 10, tr = -log10(0.05/498))
```

Index