Package 'SuperExactTest'

November 10, 2015

Type Package

Title Exact Test and Visualization of Multi-Set Intersections

Version 0.99.2

Date 2015-11-07

Author Minghui Wang, Yongzhong Zhao and Bin Zhang

Maintainer Minghui Wang <minghui.wang@mssm.edu>

Contact Minghui Wang <minghui.wang@mssm.edu>, Bin Zhang <bin.zhang@mssm.edu>

Description Efficient statistical testing and scalable visualization of intersections among multiple sets.

License GPL-3

Depends grid (>= 3.1.0), methods, R (>= 3.1.0)

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation yes

Repository CRAN

Date/Publication 2015-11-10 18:23:11

R topics documented:

`Cancer` *Cancer Census Dataset*

Description

This example dataset contains a list of seven cancer predisposition gene sets.

Usage

```
Cancer
```

Details

The seven cancer predisposition gene sets are:

1) NRG (Rahman, N. Realizing the promise of cancer predisposition genes. Nature 2014, 505:302-308);

2) NBG (Tamborero, D. et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Scientific reports 2013, 3:2650);

3) LDG (Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. Nature 2013, 502:333-339);

4) GGG (Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014, 505:495-501);

5) ELG (Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. Cell 2013, 153:17-37);

6) CCG (Futreal, P. A. et al. A census of human cancer genes. Nature reviews. Cancer 2004, 4:177-183);

7) BVG (Vogelstein, B. et al. Cancer genome landscapes. Science 2013, 339:1546-1558).

References

Minghui Wang, Yongzhong Zhao, and Bin Zhang (2015). Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports* (In press).

See Also

`supertest`

cis.eqtls *cis-eQTLs*

Description

This example dataset contains a list of cis-eQTL genes.

Details

A list is included in this dataset: `cis.eqtls`, which contains four sets of cis-eQTL genes published by Gibbs et al (PLOS Genetics 2010, 6:e1000952) as deposited in the eQTL Browser (http://www.ncbi.nlm.nih.gov/projects/g The four sets of cis-eQTL genes were detected in four different brain regions from Gibbs: brain cerebellum (CB), brain frontal cortex region (FC), brain temporal cortex region (TC), and brain pons region (PONS) respectively.

See Also

`supertest`

cpsets *Multi-Set Intersection Probability*

Description

Density and distribution function of multi-set intersection test.

Usage

```
dpsets(x,L,n,log.p =FALSE)
cpsets(x,L,n,lower.tail=TRUE,log.p=FALSE,
       simulation.p.value=FALSE,number.simulations=1000000)
```

Arguments

`x`	integer, number of elements overlap among all sets.
`L`	vector, set sizes.
`n`	integer, background population size.
`lower.tail`	logical; if TRUE, probability is `P[overlap <= x]`, otherwise, `P[overlap > x]`.
`log.p`	logical; if TRUE, probability p is given as `log(p)`.
`simulation.p.value`	logical; if TRUE, probability p is computed from simulation.
`number.simulations`	integer; number of simulations.

Value

dpsets gives the density and cpsets gives the distribution function.

Author(s)

Minghui Wang <minghui.wang@mssm.edu>, Bin Zhang <bin.zhang@mssm.edu>

References

Minghui Wang, Yongzhong Zhao, and Bin Zhang (2015). Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports* (In press).

See Also

supertest, MSET

Examples

```
##not run###
#fake data
#n=500; A=260; B=320; C=430; D=300; x=170
#(d=dpsets(x,c(A,B,C,D),n))
#(p=cpsets(x,c(A,B,C,D),n,lower.tail=FALSE))
```

GWAS *GAWS Catalog Dataset*

Description

This example dataset contains a list of gene sets associated with six types of clinical traits curated in the GWAS Catalog.

Usage

```
GWAS
```

Details

The six clinical traits are:

1) NEU (Bipolar disorder and schizophrenia, Schizophrenia, Major depressive disorder, Alzheimer's disease, Parkinson's disease, Cognitive performance, Bipolar disorder);

2) INF (Crohn's disease, Ulcerative colitis, Inflammatory bowel disease, Rheumatoid arthritis, Multiple sclerosis, Systemic lupus erythematosus);

3) CVD (Type 2 diabetes, Coronary heart disease, Blood pressure, total Cholesterol, HDL cholesterol, Triglycerides);

4) HT (height);

5) IgG (IgG glycosylation);

6) OB (obesity, obesity related traits).

References

Minghui Wang, Yongzhong Zhao, and Bin Zhang (2015). Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports* (In press).

See Also

supertest

`intersect` *Set Operations*

Description

Performs set union and intersection on multiple input vectors.

Usage

```
union(x, y, ...)
intersect(x, y, ...)
```

Arguments

`x, y, ...`	vectors (of the same mode) containing a sequence of items (conceptually) with no duplicated values.

Details

These functions extend the the same functions in the `base` package to handle more than two input vectors.

Value

A vector of the same mode as x or y for intersect, and of a common mode for union.

Author(s)

Minghui Wang <minghui.wang@mssm.edu>, Bin Zhang <bin.zhang@mssm.edu>

References

Minghui Wang, Yongzhong Zhao, and Bin Zhang (2015). Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports* (In press).

Examples

```
##not run##
```

jaccard *Calculate Jaccard Index*

Description

This function calculates Jaccard indices between pairs of sets.

Usage

```
jaccard(x)
```

Arguments

x	list, a collect of sets.

Value

A matrix of pairwise Jaccard indices.

Author(s)

Minghui Wang <minghui.wang@mssm.edu>

Examples

```
##not run###
#fake data
#x=list(S1=letters[1:20],S2=letters[10:26],S3=sample(letters,10),
# S4=sample(letters,10))
#jaccard(x)
```

MSET *Exact Test of Multi-Set Intersection*

Description

Calculate FE and significance of intersection among multiple sets.

Usage

```
MSET(x,n,lower.tail=TRUE,log.p=FALSE)
```

Arguments

x	list; a collection of sets.
n	integer; background population size.
lower.tail	logical; if TRUE, probability is P[overlap < m], otherwise, P[overlap >= m], where m is the number of elements overlap between all sets.
log.p	logical; if TRUE, probability p is given as log(p).

Value

A list with the following elements:

intersects	a vector of intersect items.
FE	fold enrichment of the intersection.
p.value	one-tail probability of observing equal to or larger than the number of intersect items.

Author(s)

Minghui Wang <minghui.wang@mssm.edu>, Bin Zhang <bin.zhang@mssm.edu>

References

Minghui Wang, Yongzhong Zhao, and Bin Zhang (2015). Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports* (In press).

See Also

supertest, cpsets, dpsets

Examples

```
##not run###
#fake data
#x=list(S1=letters[1:20],S2=letters[10:26],S3=sample(letters,10),
# S4=sample(letters,10))
#MSET(x,26,FALSE)
```

---

msets *Class to Contain Multi-Set Intersections*

Description

This object contains data regarding the intersections between multiple sets.

Details

This is an object created by the supertest function.

Value

x	a list of sets from input.
set.names	names of the sets. If the input sets do not have names, they will be automatically named as SetX where X is an integer from 1 to the total number of sets.
set.sizes	a vector of set sizes.
n	background population size.
overlap.sizes	a named vector of intersection sizes. Each intersection component is named as a character strings of 0 and 1, where a value of 1 in the ith position of the string indicates the intersection is involved with the ith set; 0 otherwise. E.g., string '000101' indicates that the intersection is an overlap between the 4th and 6th sets.
P.value	a vector of p values for the intersections.

Author(s)

Minghui Wang <minghui.wang@mssm.edu>, Bin Zhang <bin.zhang@mssm.edu>

References

Minghui Wang, Yongzhong Zhao, and Bin Zhang (2015). Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports* (In press).

See Also

supertest, summary.msets, plot.msets

---

plot.msets *Draw Multi-Set Intersections*

Description

This function draws intersections among multiple sets.

Usage

```
## S3 method for class 'msets'
plot(x, Layout=c('circular','landscape'), degree=NULL,
keep.empty.intersections=TRUE, sort.by=c('set','size','degree','p-value'),
ylim=NULL, log.scale=FALSE, x.pos=c(0.05,0.95),
y.pos=c(0.025,0.975), yfrac=0.8, color.scale.pos=c(0.85, 0.9),
legend.pos=c(0.85,0.25), legend.col=2, legend.text.cex=1, color.scale.cex=1,
color.scale.title=expression(paste(-Log[10],'(',italic(P),')')),
color.on='#2EFE64', color.off='#EEEEEE', show.overlap.size=TRUE,
show.set.size=TRUE, track.area.range=0.3, bar.area.range=0.2,
origin=if(sort.by[1]=='size'){c(0.45,0.5)}else{c(0.5,0.5)},
...)
```

Arguments

x	a msets object.
Layout	layout for plotting.
degree	a vector of intersection degrees for plotting. E.g., when degree=c(2:3), only those intersections involving two or three sets will be plotted. By default, degree=NULL, all possible intersections are plotted.
keep.empty.intersections	logical; if FALSE, empty intersection(s) will be discarded to save plotting space.
sort.by	how to sort intersections. It is one of "set", "size", "degree", and "p-value".
ylim	the limits c(y1, y2) of plotting overlap size.
log.scale	logical; whether to plot with log transformed intersection sizes.
x.pos	numeric; x coordinate (0 to 1) of the graph canvas for landscape Layout.
y.pos	numeric; y coordinate (0 to 1) of the graph canvas for landscape Layout.
yfrac	numeric; the fraction (0 to 1) of canvas used for plotting bars. Only used for landscape Layout.
color.scale.pos	numeric; x and y coordinates (0 to 1) for packing the color scale guide. It could be a keyword "topright" or "topleft" in the landscape layout, and one of "topright", "topleft", "bottomright" and "bottomleft" in the circular layout.
legend.pos	numeric; x and y coordinates (0 to 1) for packing the legend in the circular layout. It could be one of the keywords "bottomright", "bottomleft", "topleft" and "topright".
legend.col	integer; number of columns of the legend in the circular layout.
legend.text.cex	numeric; specifying the amount by which legend text should be magnified relative to the default.
color.scale.cex	numeric; specifying the amount by which color scale text should be magnified relative to the default.
color.scale.title	character or expression; a title for the color scale guide.
color.on	color code; specifying the color for set(s) which are "present" for an intersection.
color.off	color code; specifying the color for set(s) which are "absent" for an intersection.
show.overlap.size	logical; whether to show overlap size in circular layout.
show.set.size	color code; whether to show set size in landscape layout.
track.area.range	the magnitude of track area from origin in the circular layout.
bar.area.range	the magnitude of bar area from edge of the track area in the circular layout. The sum of track.area.range and track.area.range should not be larger than 0.5.

origin	the origin coordinates (0 to 1) in the circular layout.
...	additional arguments for plot function, including heatmapColor (a vector of heat colors), cex (scale of text font size), phantom.traks (number of phantom trakcs in the middle in the circular layout, default 2), gap.within.track (ratio of gap width over block width on the same track, default 0.1), and gap.between.track (ratio of gap width over track width, default 0.1). Not fully implemented.

Details

The plot canvas has coordinates 0~1 for both x and y axes.

Value

No return.

Author(s)

Minghui Wang <minghui.wang@mssm.edu>, Bin Zhang <bin.zhang@mssm.edu>

References

Minghui Wang, Yongzhong Zhao, and Bin Zhang (2015). Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports* (In press).

See Also

msets

Examples

```
##not run###
#fake data
#x=list(S1=letters[1:20],S2=letters[10:26],S3=sample(letters,10),
# S4=sample(letters,10))
#obj=supertest(x,n=26)
#plot(obj)
```

---

summary.msets *Summarize an msets Object*

Description

This function outputs summary statistics of a msets object.

Usage

```
## S3 method for class 'msets'
summary(object, degree=NULL, ...)
```

Arguments

object	a msets object.
degree	a vector of intersection degrees to pull out.
...	additional arguments (not implemented).

Value

A list:

Barcode	a vector of 0/1 character strings, representing the set composition of each intersection.
otab	a vector of observed intersection size between any combination of sets.
etab	a vector of expected intersection size between any combination of sets if background population size is specified.
set.names	set names.
set.sizes	set sizes.
n	background population size.
P.value	upper tail p value for each intersection if background population size n is specified.
Table	a data.frame containing degree, otab, etab, fold change, p value and the overlap elements.

Author(s)

Minghui Wang <minghui.wang@mssm.edu>, Bin Zhang <bin.zhang@mssm.edu>

References

Minghui Wang, Yongzhong Zhao, and Bin Zhang (2015). Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports* (In press).

See Also

msets

Examples

```
##not run###
#fake data
#x=list(S1=letters[1:20],S2=letters[10:26],S3=sample(letters,10),
# S4=sample(letters,10))
#obj=supertest(x,n=26)
#summary(obj)
```

`supertest` *Calculate Intersections Among Multiple Sets and Perform Statistical Tests*

Description

This function calculates intersection sizes among multiple sets and performs statistical tests of the intersections.

Usage

```
supertest(x, n=NULL, degree=NULL, ...)
```

Arguments

`x`	list; a collection of sets.
`n`	integer, background population size. Required for computing the statistical significance of intersections.
`degree`	a vector of intersection degrees for overlap analysis. E.g., when `degree=c(2:3)`, only those intersections involving two or three sets will be computed. By default, `degree=NULL`, all possible intersections are computed.
`...`	additional arguments (not implemented).

Details

This function calculates intersection sizes between multiple sets and, if background population size `n` is specified, performs statistical tests of the intersections.

Value

An object of class `msets`.

Author(s)

Minghui Wang <minghui.wang@mssm.edu>, Bin Zhang <bin.zhang@mssm.edu>

References

Minghui Wang, Yongzhong Zhao, and Bin Zhang (2015). Efficient Test and Visualization of Multi-Set Intersections. *Scientific Reports* (In press).

See Also

`msets`, `MSET`, `Cancer`, `cpsets`, `dpsets`

Examples

```
#Analyze the cancer gene sets
data(Cancer)
Result=supertest(Cancer, n=20687)
summary(Result)
plot(Result,degree=2:7,sort.by='size')
```

Index