Package ‘zalpha’

March 16, 2020

Type Package
Title Run a Suite of Selection Statistics
Version 0.1.0
Author Clare Horscroft
Maintainer Clare Horscroft <c.horscroft@soton.ac.uk>
License MIT + file LICENSE
Encoding UTF-8
LazyData true
RoxygenNote 7.0.2
Suggests testthat (>= 2.1.0), knitr, rmarkdown
VignetteBuilder knitr
NeedsCompilation no
Repository CRAN
Date/Publication 2020-03-16 11:50:03 UTC

R topics documented:

LDprofile ................................................................. 2
LR ................................................................. 2
L_plus_R ............................................................. 3
snpS ............................................................... 4
zalpha .............................................................. 5
zalpha_all .......................................................... 6
zalpha_BetaCDF ..................................................... 9
zalpha_expected .................................................. 11
zalpha_log_rsq_over_expected ................................ 12
zalpha_rsq_over_expected ....................................... 14
zalpha_Zscore ...................................................... 16
Index 29

LDprofile

Dataset containing an example LD profile

Description

A simulated LD profile, containing example LD statistics for genetic distances of 0 to 0.0049, in bins of size 0.0001.

Usage
data(LDprofile)

Format

A data frame with 50 rows and 5 variables:

- **bin**: the lower bound of each bin
- **rsq**: the expected $r^2$ value for a pair of SNPs, where the genetic distance between them falls in the given bin
- **sd**: the standard deviation of the expected $r^2$ value
- **Beta_a**: the first shape parameter for the Beta distribution fitted for this bin
- **Beta_b**: the second shape parameter for the Beta distribution fitted for this bin

LR

Runs the LR function

Description

Returns the $|L| |R|$ value for each SNP location supplied to the function, where $|L|$ and $|R|$ are the number of SNPs to the left and right of the current locus within the given window $ws$. For more information about the $|L| |R|$ diversity statistic, please see Jacobs (2016).

Usage

LR(pos, ws, X = NULL)
Arguments

pos  A numeric vector of SNP locations
ws  The window size which the LR statistic will be calculated over. This should be on the same scale as the pos vector.
X  Optional. Specify a region of the chromosome to calculate LR for in the format c(startposition, endposition). The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate LR for every SNP in the pos vector.

Value

A list containing the SNP positions and the LR values for those SNPs

References


Examples

## load the snps example dataset
data(snps)
## run LR over all the SNPs with a window size of 3000 bp
LR(snps$positions,3000)
## only return results for SNPs between locations 600 and 1500 bp
LR(snps$positions,3000,X=c(600,1500))

---

L_plus_R  Runs the L_plus_R function

Description

Returns the $\left\lvert L_i \right\rvert + \left\lvert R_i \right\rvert$ value for each SNP location supplied to the function. $\left\lvert L \right\rvert$ and $\left\lvert R \right\rvert$ are the number of SNPs to the left and right of the current locus within the given window ws. For more information about the L_plus_R diversity statistic, please see Jacobs (2016).

Usage

L_plus_R(pos, ws, X = NULL)
### Arguments

- **pos**
  A numeric vector of SNP locations

- **ws**
  The window size which the `L_plus_R` statistic will be calculated over. This should be on the same scale as the `pos` vector.

- **X**
  Optional. Specify a region of the chromosome to calculate `L_plus_R` for in the format `c(startposition,endposition)`. The start position and the end position should be within the extremes of the positions given in the `pos` vector. If not supplied, the function will calculate `L_plus_R` for every SNP in the `pos` vector.

### Value

A list containing the SNP positions and the `L_plus_R` values for those SNPs

### References


### Examples

```r
## load the snps example dataset
data(snps)
## run `L_plus_R` over all the SNPs with a window size of 3000 bp
L_plus_R(snps$positions,3000)
## only return results for SNPs between locations 600 and 1500 bp
L_plus_R(snps$positions,3000,X=c(600,1500))
```

---

<table>
<thead>
<tr>
<th>snps</th>
<th>Dataset containing details on simulated SNPs</th>
</tr>
</thead>
</table>

### Description

A dataset containing the positions, genetic distances and alleles for 20 SNPs, across 10 simulated chromosomes.

### Usage

```r
snps
```
Zalpha

Format
A data frame with 20 rows and 12 variables:

positions location of the SNP on the chromosome
distances genetic distance of the SNP from the start of the chromosome
c Chrom_1 allele of the SNP on the first example chromosome
c Chrom_2 allele of the SNP on the second example chromosome
c Chrom_3 allele of the SNP on the third example chromosome
c Chrom_4 allele of the SNP on the fourth example chromosome
c Chrom_5 allele of the SNP on the fifth example chromosome
c Chrom_6 allele of the SNP on the sixth example chromosome
c Chrom_7 allele of the SNP on the seventh example chromosome
c Chrom_8 allele of the SNP on the eighth example chromosome
c Chrom_9 allele of the SNP on the ninth example chromosome
c Chrom_10 allele of the SNP on the tenth example chromosome

Examples

snps

Zalpha

Runs the Zalpha function

Description
Returns a $Z_\alpha$ value for each SNP location supplied to the function. For more information about the $Z_\alpha$ statistic, please see Jacobs (2016). The $Z_\alpha$ statistic is defined as:

$$Z_\alpha = \left(\frac{|L|}{2}\right)^{-1} \sum_{i,j \in L} r_{i,j}^2 + \left(\frac{|R|}{2}\right)^{-1} \sum_{i,j \in L} r_{i,j}^2$$

where $|L|$ and $|R|$ are the number of SNPs to the left and right of the current locus within the given window $ws$, and $r^2$ is equal to the squared correlation between a pair of SNPs

Usage
Zalpha(pos, ws, x, minRandL = 4, minRL = 25, X = NULL)
Arguments

- **pos**: A numeric vector of SNP locations
- **ws**: The window size which the Zα statistic will be calculated over. This should be on the same scale as the pos vector.
- **x**: A matrix of SNP values. Columns represent chromosomes; rows are SNP locations. Hence, the number of rows should equal the length of the pos vector. SNPs should all be biallelic.
- **minRandL**: Minimum number of SNPs in each set R and L for the statistic to be calculated. Default is 4.
- **minRL**: Minimum value for the product of the set sizes for R and L. Default is 25.
- **X**: Optional. Specify a region of the chromosome to calculate Zα for in the format `c(startposition,endposition)`. The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate Zα for every SNP in the pos vector.

Value

A list containing the SNP positions and the Zα values for those SNPs

References


Examples

```r
## load the snps example dataset
data(snps)
## run Zalpha over all the SNPs with a window size of 3000 bp
Zalpha(snps$positions, 3000, as.matrix(snps[,3:12]))
## only return results for SNPs between locations 600 and 1500 bp
Zalpha(snps$positions, 3000, as.matrix(snps[,3:12]), X = c(600, 1500))
```

---

**Zalpha_all**  
*Runs all the statistics in the zalpha package*

Description

Returns every statistic for each SNP location, given the appropriate parameters. See Details for more information.
Usage

Zalpha_all(
  pos,
  ws,
  x = NULL,
  dist = NULL,
  LDprofile_bins = NULL,
  LDprofile_rsq = NULL,
  LDprofile_sd = NULL,
  LDprofile_Beta_a = NULL,
  LDprofile_Beta_b = NULL,
  minRandL = 4,
  minRL = 25,
  X = NULL
)

Arguments

pos
  A numeric vector of SNP locations

ws
  The window size which the statistics will be calculated over. This should be on the same scale as the pos vector.

x
  Optional. A matrix of SNP values. Columns represent chromosomes; rows are SNP locations. Hence, the number of rows should equal the length of the pos vector. SNPs should all be biallelic.

dist
  Optional. A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as pos.

LDprofile_bins
  Optional. A numeric vector containing the lower bound of the bins used in the LD profile. These should be of equal size.

LDprofile_rsq
  Optional. A numeric vector containing the expected $r^2$ values for the corresponding bin in the LD profile. Must be between 0 and 1.

LDprofile_sd
  Optional. A numeric vector containing the standard deviation of the $r^2$ values for the corresponding bin in the LD profile.

LDprofile_Beta_a
  Optional. A numeric vector containing the first estimated Beta parameter for the corresponding bin in the LD profile.

LDprofile_Beta_b
  Optional. A numeric vector containing the second estimated Beta parameter for the corresponding bin in the LD profile.

minRandL
  Minimum number of SNPs in each set R and L for the statistics to be calculated. Default is 4.

minRL
  Minimum value for the product of the set sizes for R and L. Default is 25.

X
  Optional. Specify a region of the chromosome to calculate the statistics for in the format c(startposition,endposition). The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate the statistics for every SNP in the pos vector.
Details

Not all statistics will be returned, depending on the parameters supplied to the function. If \( x \) is not supplied, only \( \text{Zalpha\_expected}, \text{Zbeta\_expected}, \text{LR} \) and \( \text{L\_plus\_R} \) will be calculated. For any of the statistics which use an expected \( r^2 \) value, the parameters \text{dist}, \text{LDprofile\_bins} \) and \text{LDprofile\_rsq} must be supplied. This includes the statistics: \( \text{Zalpha\_expected}, \text{Zalpha\_rsq\_over\_expected}, \text{Zalpha\_log\_rsq\_over\_expected}, \text{Zalpha\_Zscore}, \text{Zalpha\_BetaCDF}, \text{Zbeta\_expected}, \text{Zbeta\_rsq\_over\_expected}, \text{Zbeta\_log\_rsq\_over\_expected}, \text{Zbeta\_Zscore} \) and \( \text{Zbeta\_BetaCDF} \).

- For \( \text{Zalpha\_Zscore} \) and \( \text{Zbeta\_Zscore} \) to be calculated, the parameter \text{LDprofile\_sd} must also be supplied.
- For \( \text{Zalpha\_BetaCDF} \) and \( \text{Zbeta\_BetaCDF} \) to be calculated, the parameters \text{LDprofile\_Beta\_a} \) and \text{LDprofile\_Beta\_b} must also be supplied.

For more information about the statistics, please see Jacobs (2016).

Value

A list containing the SNP positions and the statistics for those SNPs

References


See Also

\( \text{Zalpha\_Zalpha\_expected}, \text{Zalpha\_rsq\_over\_expected}, \text{Zalpha\_log\_rsq\_over\_expected}, \text{Zalpha\_Zscore}, \text{Zalpha\_BetaCDF}, \text{Zbeta\_Zbeta\_expected}, \text{Zbeta\_rsq\_over\_expected}, \text{Zbeta\_log\_rsq\_over\_expected}, \text{Zbeta\_Zscore}, \text{Zbeta\_BetaCDF}, \text{LR}, \text{L\_plus\_R} \)

Examples

```R
## load the snps and LDprofile example datasets
data(snps)
data(LDprofile)
## run Zalpha\_all over all the SNPs with a window size of 3000 bp
## will return all 15 statistics
Zalpha\_all(snps\$positions, 3000, as.matrix(snps[, 3:12]), snps\$distances, LDprofile\$bin, LDprofile\$rsq, LDprofile\$sd, LDprofile\$Beta\_a, LDprofile\$Beta\_b)
## only return results for SNPs between locations 600 and 1500 bp
Zalpha\_all(snps\$positions, 3000, as.matrix(snps[, 3:12]), snps\$distances, LDprofile\$bin, LDprofile\$rsq, LDprofile\$sd, LDprofile\$Beta\_a, LDprofile\$Beta\_b, X = c(600, 1500))
## will only return statistics not requiring an LD profile
Zalpha\_all(snps\$positions, 3000, as.matrix(snps[, 3:12]))
```
Zalpha_BetaCDF runs the Zalpha function using a cumulative beta distribution function on the r-squared values for the region

### Description

Returns a \( Z_{\alpha BetaCDF} \) value for each SNP location supplied to the function, based on the expected \( r^2 \) values given an LD profile and genetic distances. For more information about the \( Z_{\alpha BetaCDF} \) statistic, please see Jacobs (2016). The \( Z_{\alpha BetaCDF} \) statistic is defined as:

\[
Z_{\alpha BetaCDF} = \frac{\left(\frac{|L|}{2}\right)^{-1} \sum_{i,j \in L} B(r_{i,j}^2; a,b) + \left(\frac{|R|}{2}\right)^{-1} \sum_{i,j \in R} B(r_{i,j}^2; a,b)}{2}
\]

where \(|L|\) and \(|R|\) are the number of SNPs to the left and right of the current locus within the given window \( ws \), \( r^2 \) is equal to the squared correlation between a pair of SNPs, and \( \frac{B(r_{i,j}^2; a,b)}{B(a,b)} \) is the cumulative distribution function for the Beta distribution given the estimated \( a \) and \( b \) parameters from the LD profile.

### Usage

```r
Zalpha_BetaCDF(
  pos,
  ws,
  x,
  dist,
  LDprofile_bins,
  LDprofile_Beta_a,
  LDprofile_Beta_b,
  minRandL = 4,
  minRL = 25,
  X = NULL
)
```

### Arguments

- **pos**: A numeric vector of SNP locations
- **ws**: The window size which the \( Z_{\alpha BetaCDF} \) statistic will be calculated over. This should be on the same scale as the pos vector.
- **x**: A matrix of SNP values. Columns represent chromosomes; rows are SNP locations. Hence, the number of rows should equal the length of the pos vector. SNPs should all be biallelic.
- **dist**: A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as pos.
- **LDprofile_bins**: A numeric vector containing the lower bound of the bins used in the LD profile. These should be of equal size.
LDprofile_Beta_a
A numeric vector containing the first estimated Beta parameter for the corresponding bin in the LD profile.

LDprofile_Beta_b
A numeric vector containing the second estimated Beta parameter for the corresponding bin in the LD profile.

minRandL Minimum number of SNPs in each set R and L for the statistic to be calculated. Default is 4.

minRL Minimum value for the product of the set sizes for R and L. Default is 25.

X Optional. Specify a region of the chromosome to calculate $Z_{\alpha}^{BetaCDF}$ for in the format c(startposition,endposition). The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate $Z_{\alpha}^{BetaCDF}$ for every SNP in the pos vector.

Details

The LD profile describes the expected correlation between SNPs at a given genetic distance, generated using simulations or real data. Care should be taken to utilise an LD profile that is representative of the population in question. The LD profile should consist of evenly sized bins of distances (for example 0.0001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

Value

A list containing the SNP positions and the $Z_{\alpha}^{BetaCDF}$ values for those SNPs

References


Examples

```r
## load the snps and LDprofile example datasets
data(snps)
data(LDprofile)
## run Zalpha_BetaCDF over all the SNPs with a window size of 3000 bp
Zalpha_BetaCDF(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances,
LDprofile$bin,LDprofile$Beta_a,LDprofile$Beta_b)
## only return results for SNPs between locations 600 and 1500 bp
Zalpha_BetaCDF(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances,
LDprofile$bin,LDprofile$Beta_a,LDprofile$Beta_b,X=c(600,1500))
```
Zalpha_expected

**Description**

Runs the Zalpha function on the expected r-squared values for the region

Returns a $Z_{E[r^2]}$ value for each SNP location supplied to the function, based on the expected $r^2$ values given an LD profile and genetic distances. For more information about the $Z_{E[r^2]}$ statistic, please see Jacobs (2016). The $Z_{E[r^2]}$ statistic is defined as:

$$Z_{E[r^2]} = \left( \frac{|L|}{2} \right)^{-1} \sum_{i,j \in L} E[r_{i,j}^2] + \left( \frac{|R|}{2} \right)^{-1} \sum_{i,j \in R} E[r_{i,j}^2]$$

where $|L|$ and $|R|$ are the number of SNPs to the left and right of the current locus within the given window $ws$, and $E[r^2]$ is equal to the expected squared correlation between a pair of SNPs, given an LD profile.

**Usage**

```r
Zalpha_expected(
  pos, 
  ws, 
  dist, 
  LDprofile_bins, 
  LDprofile_rsq, 
  minRandL = 4, 
  minRL = 25, 
  X = NULL 
)
```

**Arguments**

- `pos`: A numeric vector of SNP locations
- `ws`: The window size which the $Z_{E[r^2]}$ statistic will be calculated over. This should be on the same scale as the `pos` vector.
- `dist`: A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as `pos`.
- `LDprofile_bins`: A numeric vector containing the lower bound of the bins used in the LD profile. These should be of equal size.
- `LDprofile_rsq`: A numeric vector containing the expected $r^2$ values for the corresponding bin in the LD profile. Must be between 0 and 1.
- `minRandL`: Minimum number of SNPs in each set R and L for the statistic to be calculated. Default is 4.
- `minRL`: Minimum value for the product of the set sizes for R and L. Default is 25.
Zalpha_log_rsq_over_expected

X Optional. Specify a region of the chromosome to calculate $Z_{\alpha}^{E[r^2]}$ for in the format c(startposition,endposition). The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate $Z_{\alpha}^{E[r^2]}$ for every SNP in the pos vector.

Details

The LD profile describes the expected correlation between SNPs at a given genetic distance, generated using simulations or real data. Care should be taken to utilise an LD profile that is representative of the population in question. The LD profile should consist of evenly sized bins of distances (for example 0.0001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

Value

A list containing the SNP positions and the $Z_{\alpha}^{E[r^2]}$ values for those SNPs

References


Examples

```r
## load the snps and LDprofile example datasets
data(snps)
data(LDprofile)
## run Zalpha_expected over all the SNPs with a window size of 3000 bp
Zalpha_expected(snps$positions,3000,snps$distances,LDprofile$bin,LDprofile$rsq)
## only return results for SNPs between locations 600 and 1500 bp
Zalpha_expected(snps$positions,3000,snps$distances,LDprofile$bin,LDprofile$rsq,X=c(600,1500))
```

---

Zalpha_log_rsq_over_expected

Runs the Zalpha function on the log of the r-squared values over the expected r-squared values for the region

Description

Returns a $Z_{\alpha}^{\log_{10}(r^2/E[r^2])}$ value for each SNP location supplied to the function, based on the expected $r^2$ values given an LD profile and genetic distances. For more information about the $Z_{\alpha}^{\log_{10}(r^2/E[r^2])}$ statistic, please see Jacobs (2016). The $Z_{\alpha}^{\log_{10}(r^2/E[r^2])}$ statistic is defined as:

$$Z_{\alpha}^{\log_{10}(r^2/E[r^2])} = \frac{\frac{|L|}{2} - 1}{2} \sum_{i,j \in L} \log_{10}(r_{i,j}^2/E[r_{i,j}^2]) + \frac{\frac{|R|}{2} - 1}{2} \sum_{i,j \in R} \log_{10}(r_{i,j}^2/E[r_{i,j}^2])$$

where $|L|$ and $|R|$ are the number of SNPs to the left and right of the current locus within the given window $ws$, $r^2$ is equal to the squared correlation between a pair of SNPs, and $E[r^2]$ is equal to the expected squared correlation between a pair of SNPs, given an LD profile.
Usage

\texttt{Zalpha_log_rsq_over_expected(pos, ws, x, dist, LDprofile_bins, LDprofile_rsq, minRandL = 4, minRL = 25, X = NULL)}

Arguments

\begin{itemize}
\item \texttt{pos} A numeric vector of SNP locations
\item \texttt{ws} The window size which the $Z_{\alpha}^{log_{10}(r^2/E[r^2])}$ statistic will be calculated over. This should be on the same scale as the pos vector.
\item \texttt{x} A matrix of SNP values. Columns represent chromosomes; rows are SNP locations. Hence, the number of rows should equal the length of the pos vector. SNPs should all be biallelic.
\item \texttt{dist} A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as pos.
\item \texttt{LDprofile_bins} A numeric vector containing the lower bound of the bins used in the LD profile. These should be of equal size.
\item \texttt{LDprofile_rsq} A numeric vector containing the expected $r^2$ values for the corresponding bin in the LD profile. Must be between 0 and 1.
\item \texttt{minRandL} Minimum number of SNPs in each set R and L for the statistic to be calculated. Default is 4.
\item \texttt{minRL} Minimum value for the product of the set sizes for R and L. Default is 25.
\item \texttt{X} Optional. Specify a region of the chromosome to calculate $Z_{\alpha}^{log_{10}(r^2/E[r^2])}$ for in the format \texttt{c(startposition,endposition)}. The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate $Z_{\alpha}^{log_{10}(r^2/E[r^2])}$ for every SNP in the pos vector.
\end{itemize}

Details

The LD profile describes the expected correlation between SNPs at a given genetic distance, generated using simulations or real data. Care should be taken to utilise an LD profile that is representative of the population in question. The LD profile should consist of evenly sized bins of distances (for example 0.0001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

Value

A list containing the SNP positions and the $Z_{\alpha}^{log_{10}(r^2/E[r^2])}$ values for those SNPs
References


Examples

```r
## load the snps and LDprofile example datasets
data(snps)
data(LDprofile)
## run Zalpha_log_rsq_over_expected over all the SNPs with a window size of 3000 bp
Zalpha_log_rsq_over_expected(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances,
LDprofile$bin,LDprofile$rsq)
## only return results for SNPs between locations 600 and 1500 bp
Zalpha_log_rsq_over_expected(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances,
LDprofile$bin,LDprofile$rsq,X=c(600,1500))
```

---

**Zalpha_rsq_over_expected**

*Runs the Zalpha function on the r-squared values over the expected r-squared values for the region*

**Description**

Returns a \( Z_{\alpha}^{r^2/E[r^2]} \) value for each SNP location supplied to the function, based on the expected \( r^2 \) values given an LD profile and genetic distances. For more information about the \( Z_{\alpha}^{r^2/E[r^2]} \) statistic, please see Jacobs (2016). The \( Z_{\alpha}^{r^2/E[r^2]} \) statistic is defined as:

\[
Z_{\alpha}^{r^2/E[r^2]} = \left( \frac{|L|}{2} \right)^{-1} \sum_{i,j \in L} \frac{r^2_{i,j}}{E[r^2_{i,j}]} + \left( \frac{|R|}{2} \right)^{-1} \sum_{i,j \in R} \frac{r^2_{i,j}}{E[r^2_{i,j}]}
\]

where \(|L|\) and \(|R|\) are the number of SNPs to the left and right of the current locus within the given window \( ws \), \( r^2 \) is equal to the squared correlation between a pair of SNPs, and \( E[r^2] \) is equal to the expected squared correlation between a pair of SNPs, given an LD profile.

**Usage**

```r
Zalpha_rsq_over_expected(
    pos,  # SNP positions
    ws,   # window size
    x,    # data frame with r^2 values
    dist, # genetic distances
    LDprofile_bins,  # bins from LD profile
    LDprofile_rsq,   # r^2 values from LD profile
    minRandL = 4,    # minimum random SNPs to the left
    minRL = 25,      # minimum random SNPs to the right
    X = NULL         # optional start and end positions
)
```
Arguments

pos          A numeric vector of SNP locations
ws           The window size which the $Z_{\alpha}^{r^2}/E[r^2]$ statistic will be calculated over. This should be on the same scale as the pos vector.
x            A matrix of SNP values. Columns represent chromosomes; rows are SNP locations. Hence, the number of rows should equal the length of the pos vector. SNPs should all be biallelic.
dist         A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as pos.
LDprofile_bins A numeric vector containing the lower bound of the bins used in the LD profile. These should be of equal size.
LDprofile_rsq A numeric vector containing the expected $r^2$ values for the corresponding bin in the LD profile. Must be between 0 and 1.
minRandL     Minimum number of SNPs in each set R and L for the statistic to be calculated. Default is 4.
minRL        Minimum value for the product of the set sizes for R and L. Default is 25.
x            Optional. Specify a region of the chromosome to calculate $Z_{\alpha}^{r^2}/E[r^2]$ for in the format c(startposition,endposition). The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate $Z_{\alpha}^{r^2}/E[r^2]$ for every SNP in the pos vector.

Details

The LD profile describes the expected correlation between SNPs at a given genetic distance, generated using simulations or real data. Care should be taken to utilise an LD profile that is representative of the population in question. The LD profile should consist of evenly sized bins of distances (for example 0.0001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

Value

A list containing the SNP positions and the $Z_{\alpha}^{r^2}/E[r^2]$ values for those SNPs

References


Examples

```r
## load the snps and LDprofile example datasets
data(snps)
data(LDprofile)
## run Zalpha_rsq_over_expected over all the SNPs with a window size of 3000 bp
Zalpha_rsq_over_expected(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances, LDprofile$bin,LDprofile$rsq)
```
Zalpha_Zscore

Runs the Zalpha function using the Z score of the r-squared values for the region

### Description

Returns a \( Z_{\alpha}^{Z_{\text{score}}} \) value for each SNP location supplied to the function, based on the expected \( r^2 \) values given an LD profile and genetic distances. For more information about the \( Z_{\alpha}^{Z_{\text{score}}} \) statistic, please see Jacobs (2016). The \( Z_{\alpha}^{Z_{\text{score}}} \) statistic is defined as:

\[
Z_{\alpha}^{Z_{\text{score}}} = \frac{\left( |L| \right)^{-1} \sum_{i,j \in L} \frac{r_{i,j}^2 - E[r_{i,j}^2]}{\sigma[r_{i,j}^2]} + \left( |R| \right)^{-1} \sum_{i,j \in R} \frac{r_{i,j}^2 - E[r_{i,j}^2]}{\sigma[r_{i,j}^2]}}{2}
\]

where \(|L|\) and \(|R|\) are the number of SNPs to the left and right of the current locus within the given window \( ws \), \( r^2 \) is equal to the squared correlation between a pair of SNPs, \( E[r^2] \) is equal to the expected squared correlation between a pair of SNPs, given an LD profile, and \( \sigma[r^2] \) is the standard deviation.

### Usage

```
Zalpha_Zscore(pos, ws, x, dist, LDprofile_bins, LDprofile_rsq, LDprofile_sd, minRandL = 4, minRL = 25, X = NULL)
```

### Arguments

- **pos**
  A numeric vector of SNP locations

- **ws**
  The window size which the \( Z_{\alpha}^{Z_{\text{score}}} \) statistic will be calculated over. This should be on the same scale as the pos vector.

- **x**
  A matrix of SNP values. Columns represent chromosomes; rows are SNP locations. Hence, the number of rows should equal the length of the pos vector. SNPs should all be biallelic.
dist A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as pos.

LDprofile_bins A numeric vector containing the lower bound of the bins used in the LD profile. These should be of equal size.

LDprofile_rsq A numeric vector containing the expected $r^2$ values for the corresponding bin in the LD profile. Must be between 0 and 1.

LDprofile_sd A numeric vector containing the standard deviation of the $r^2$ values for the corresponding bin in the LD profile.

minRandL Minimum number of SNPs in each set R and L for the statistic to be calculated. Default is 4.

minRL Minimum value for the product of the set sizes for R and L. Default is 25.

X Optional. Specify a region of the chromosome to calculate $Z_{\alpha}^{Zscore}$ for in the format c(startposition,endposition). The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate $Z_{\alpha}^{Zscore}$ for every SNP in the pos vector.

Details

The LD profile describes the expected correlation between SNPs at a given genetic distance, generated using simulations or real data. Care should be taken to utilise an LD profile that is representative of the population in question. The LD profile should consist of evenly sized bins of distances (for example 0.0001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

Value

A list containing the SNP positions and the $Z_{\alpha}^{Zscore}$ values for those SNPs

References


Examples

```r
## load the snps and LDprofile example datasets
data(snps)
data(LDprofile)
## run Zalpha_Zscore over all the SNPs with a window size of 3000 bp
Zalpha_Zscore(snps$positions, 3000, as.matrix(snps[, 3:12]), snps$distances,
             LDprofile$bin, LDprofile$rsq, LDprofile$sd)
## only return results for SNPs between locations 600 and 1500 bp
Zalpha_Zscore(snps$positions, 3000, as.matrix(snps[, 3:12]), snps$distances,
             LDprofile$bin, LDprofile$rsq, LDprofile$sd, X = c(600, 1500))
```
Zbeta runs the Zbeta function

Description

Returns a $Z_{\beta}$ value for each SNP location supplied to the function. For more information about the $Z_{\beta}$ statistic, please see Jacobs (2016). The $Z_{\beta}$ statistic is defined as:

$$Z_{\beta} = \sum_{i \in L, j \in R} r_{i,j}^2 \frac{|L||R|}{|L||R|}$$

where $|L|$ and $|R|$ are the number of SNPs to the left and right of the current locus within the given window $ws$, and $r^2$ is equal to the squared correlation between a pair of SNPs.

Usage

Zbeta(pos, ws, x, minRandL = 4, minRL = 25, X = NULL)

Arguments

pos A numeric vector of SNP locations
ws The window size which the $Z_{\beta}$ statistic will be calculated over. This should be on the same scale as the pos vector.
x A matrix of SNP values. Columns represent chromosomes; rows are SNP locations. Hence, the number of rows should equal the length of the pos vector. SNPs should all be biallelic.
minRandL Minimum number of SNPs in each set R and L for the statistic to be calculated. Default is 4.
minRL Minimum value for the product of the set sizes for R and L. Default is 25.
X Optional. Specify a region of the chromosome to calculate $Z_{\beta}$ for in the format c(startposition, endposition). The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate $Z_{\beta}$ for every SNP in the pos vector.

Value

A list containing the SNP positions and the $Z_{\beta}$ values for those SNPs

References

### Examples

```r
## load the snps example dataset
data(snps)
## run Zbeta over all the SNPs with a window size of 3000 bp
Zbeta(snps$positions,3000,as.matrix(snps[,3:12]))
## only return results for SNPs between locations 600 and 1500 bp
Zbeta(snps$positions,3000,as.matrix(snps[,3:12]),X=c(600,1500))
```

---

**Zbeta_BetaCDF**

Runs the Zbeta function using a cumulative beta distribution function on the \( r^2 \) squared values for the region

---

**Description**

Returns a \( Z_{\beta_{\text{BetaCDF}}} \) value for each SNP location supplied to the function, based on the expected \( r^2 \) values given an LD profile and genetic distances. For more information about the \( Z_{\beta_{\text{BetaCDF}}} \) statistic, please see Jacobs (2016). The \( Z_{\beta_{\text{BetaCDF}}} \) statistic is defined as:

\[
Z_{\beta_{\text{BetaCDF}}} = \frac{\sum_{i \in L, j \in R} B(r^2_{i,j};a,b)}{|L||R|} B(a,b)
\]

where \(|L| \) and \(|R| \) are the number of SNPs to the left and right of the current locus within the given window \( ws \), \( r^2 \) is equal to the squared correlation between a pair of SNPs, and \( \frac{B(r^2_{i,j};a,b)}{B(a,b)} \) is the cumulative distribution function for the Beta distribution given the estimated \( a \) and \( b \) parameters from the LD profile.

**Usage**

```r
Zbeta_BetaCDF(
  pos, ws, x, dist, LDprofile_bins, LDprofile_Beta_a, LDprofile_Beta_b, minRandL = 4, minRL = 25, X = NULL
)
```

**Arguments**

- `pos`: A numeric vector of SNP locations
- `ws`: The window size which the \( Z_{\beta_{\text{BetaCDF}}} \) statistic will be calculated over. This should be on the same scale as the `pos` vector.
A matrix of SNP values. Columns represent chromosomes; rows are SNP locations. Hence, the number of rows should equal the length of the pos vector. SNPs should all be biallelic.

A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as pos.

A numeric vector containing the lower bound of the bins used in the LD profile. These should be of equal size.

A numeric vector containing the first estimated Beta parameter for the corresponding bin in the LD profile.

A numeric vector containing the second estimated Beta parameter for the corresponding bin in the LD profile.

Minimum number of SNPs in each set R and L for the statistic to be calculated. Default is 4.

Minimum value for the product of the set sizes for R and L. Default is 25.

Optional. Specify a region of the chromosome to calculate $Z_{\beta}^{BetaCDF}$ for in the format c(startposition,endposition). The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate $Z_{\beta}^{BetaCDF}$ for every SNP in the pos vector.

The LD profile describes the expected correlation between SNPs at a given genetic distance, generated using simulations or real data. Care should be taken to utilise an LD profile that is representative of the population in question. The LD profile should consist of evenly sized bins of distances (for example 0.0001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

A list containing the SNP positions and the $Z_{\beta}^{BetaCDF}$ values for those SNPs


## load the snps and LDprofile example datasets
data(snps)
data(LDprofile)
## run Zbeta_BetaCDF over all the SNPs with a window size of 3000 bp
Zbeta_BetaCDF(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances, LDprofile$bin,LDprofile$Beta_a,LDprofile$Beta_b)
## only return results for SNPs between locations 600 and 1500 bp
Zbeta_BetaCDF(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances,
Zbeta_expected

Runs the Zbeta function on the expected r-squared values for the region

Description

Returns a $Z^E_{\beta r^2}$ value for each SNP location supplied to the function, based on the expected $r^2$ values given an LD profile and genetic distances. For more information about the $Z^E_{\beta r^2}$ statistic, please see Jacobs (2016). The $Z^E_{\beta r^2}$ statistic is defined as:

$$Z^E_{\beta r^2} = \frac{\sum_{i \in L, j \in R} E[r^2_{i,j}]}{|L||R|}$$

where $|L|$ and $|R|$ are the number of SNPs to the left and right of the current locus within the given window ws, and $E[r^2]$ is equal to the expected squared correlation between a pair of SNPs, given an LD profile.

Usage

Zbeta_expected(
  pos,
  ws,
  dist,
  LDprofile_bins,
  LDprofile_rsq,
  minRandL = 4,
  minRL = 25,
  X = NULL
)

Arguments

pos A numeric vector of SNP locations
ws The window size which the $Z^E_{\beta r^2}$ statistic will be calculated over. This should be on the same scale as the pos vector.
dist A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as pos.
LDprofile_bins A numeric vector containing the lower bound of the bins used in the LD profile. These should be of equal size.
LDprofile_rsq A numeric vector containing the expected $r^2$ values for the corresponding bin in the LD profile. Must be between 0 and 1.
Zbeta_log_rsq_over_expected

**minRandL**  
Minimum number of SNPs in each set R and L for the statistic to be calculated. Default is 4.

**minRL**  
Minimum value for the product of the set sizes for R and L. Default is 25.

**X**  
Optional. Specify a region of the chromosome to calculate \( Z^\beta_{\log r^2/E[r^2]} \) for in the format c(startposition,endposition). The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate \( Z^\beta_{\log r^2/E[r^2]} \) for every SNP in the pos vector.

**Details**

The LD profile describes the expected correlation between SNPs at a given genetic distance, generated using simulations or real data. Care should be taken to utilise an LD profile that is representative of the population in question. The LD profile should consist of evenly sized bins of distances (for example 0.0001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

**Value**

A list containing the SNP positions and the \( Z^\beta_{\log r^2/E[r^2]} \) values for those SNPs

**References**


**Examples**

```r
## load the snps and LDprofile example datasets
data(snps)
data(LDprofile)
## run Zbeta_expected over all the SNPs with a window size of 3000 bp
Zbeta_expected(snps$positions,3000,snps$distances,LDprofile$bin,LDprofile$rsq)
## only return results for SNPs between locations 600 and 1500 bp
Zbeta_expected(snps$positions,3000,snps$distances,LDprofile$bin,LDprofile$rsq,X=c(600,1500))
```

---

\[
Z^\beta_{\log r^2/E[r^2]} = \sum_{i \in R, j \in L} \frac{\log_{10} (r_{i,j}^2/E[r_{i,j}^2])}{|L||R|}
\]

**Description**

Returns a \( Z^\beta_{\log r^2/E[r^2]} \) value for each SNP location supplied to the function, based on the expected \( r^2 \) values given an LD profile and genetic distances. For more information about the \( Z^\beta_{\log r^2/E[r^2]} \) statistic, please see Jacobs (2016). The \( Z^\beta_{\log r^2/E[r^2]} \) statistic is defined as:
where \(|L|\) and \(|R|\) are the number of SNPs to the left and right of the current locus within the given window \(ws\), \(r^2\) is equal to the squared correlation between a pair of SNPs, and \(E[r^2]\) is equal to the expected squared correlation between a pair of SNPs, given an LD profile.

Usage

\[
Z_{\beta} \log_{10}(r^2/E[r^2]) \]

Arguments

- **pos**: A numeric vector of SNP locations
- **ws**: The window size which the \(Z_{\beta} \log_{10}(r^2/E[r^2])\) statistic will be calculated over. This should be on the same scale as the \(pos\) vector.
- **x**: A matrix of SNP values. Columns represent chromosomes; rows are SNP locations. Hence, the number of rows should equal the length of the \(pos\) vector. SNPs should all be biallelic.
- **dist**: A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as \(pos\).
- **LDprofile_bins**: A numeric vector containing the lower bound of the bins used in the LD profile. These should be of equal size.
- **LDprofile_rsq**: A numeric vector containing the expected \(r^2\) values for the corresponding bin in the LD profile. Must be between 0 and 1.
- **minRandL**: Minimum number of SNPs in each set \(R\) and \(L\) for the statistic to be calculated. Default is 4.
- **minRL**: Minimum value for the product of the set sizes for \(R\) and \(L\). Default is 25.
- **X**: Optional. Specify a region of the chromosome to calculate \(Z_{\beta} \log_{10}(r^2/E[r^2])\) for in the format \(c(\text{startposition}, \text{endposition})\). The start position and the end position should be within the extremes of the positions given in the \(pos\) vector. If not supplied, the function will calculate \(Z_{\beta} \log_{10}(r^2/E[r^2])\) for every SNP in the \(pos\) vector.

Details

The LD profile describes the expected correlation between SNPs at a given genetic distance, generated using simulations or real data. Care should be taken to utilise an LD profile that is representative of the population in question. The LD profile should consist of evenly sized bins of distances (for example 0.0001 cM per bin), where the value given is the (inclusive) lower bound of the bin.
Zbeta_rsq_over_expected

Value

A list containing the SNP positions and the $Z^{\log_{10}(r^2/E[r^2])}$ values for those SNPs

References


Examples

```r
## load the snps and LDprofile example datasets
data(snps)
data(LDprofile)
## run Zbeta_log_rsq_over_expected over all the SNPs with a window size of 3000 bp
Zbeta_log_rsq_over_expected(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances,
LDprofile$bin,LDprofile$rsq)
## only return results for SNPs between locations 600 and 1500 bp
Zbeta_log_rsq_over_expected(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances,
LDprofile$bin,LDprofile$rsq,X=c(600,1500))
```

Zbeta_rsq_over_expected

* Runs the Zbeta function on the r-squared values over the expected r-squared values for the region *

Description

Returns a $Z^{r^2/E[r^2]}$ value for each SNP location supplied to the function, based on the expected $r^2$ values given an LD profile and genetic distances. For more information about the $Z^{r^2/E[r^2]}$ statistic, please see Jacobs (2016). The $Z^{r^2/E[r^2]}$ statistic is defined as:

$$Z^{r^2/E[r^2]} = \frac{\sum_{i \in L,j \in R} r_{i,j}^2 / E[r_{i,j}^2]}{|L||R|}$$

where |L| and |R| are the number of SNPs to the left and right of the current locus within the given window ws, $r^2$ is equal to the squared correlation between a pair of SNPs, and $E[r^2]$ is equal to the expected squared correlation between a pair of SNPs, given an LD profile.

Usage

```r
Zbeta_rsq_over_expected(
  pos,
  ws,
  x,
  dist,
  LDprofile_bins,
```

```
\[
\text{Zbeta_rsq_over_expected}
\]

\[
\text{LDprofile_rsq,}
\text{minRandL = 4,}
\text{minRL = 25,}
\text{X = NULL}
\]

\)

**Arguments**

- **pos**: A numeric vector of SNP locations
- **ws**: The window size which the \(Z_{\beta} r^2/E[r^2]\) statistic will be calculated over. This should be on the same scale as the pos vector.
- **x**: A matrix of SNP values. Columns represent chromosomes; rows are SNP locations. Hence, the number of rows should equal the length of the pos vector. SNPs should all be biallelic.
- **dist**: A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as pos.
- **LDprofile_bins**: A numeric vector containing the lower bound of the bins used in the LD profile. These should be of equal size.
- **LDprofile_rsq**: A numeric vector containing the expected \(r^2\) values for the corresponding bin in the LD profile. Must be between 0 and 1.
- **minRandL**: Minimum number of SNPs in each set R and L for the statistic to be calculated. Default is 4.
- **minRL**: Minimum value for the product of the set sizes for R and L. Default is 25.
- **X**: Optional. Specify a region of the chromosome to calculate \(Z_{\beta} r^2/E[r^2]\) for in the format \(c(\text{startposition}, \text{endposition})\). The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate \(Z_{\beta} r^2/E[r^2]\) for every SNP in the pos vector.

**Details**

The LD profile describes the expected correlation between SNPs at a given genetic distance, generated using simulations or real data. Care should be taken to utilise an LD profile that is representative of the population in question. The LD profile should consist of evenly sized bins of distances (for example 0.0001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

**Value**

A list containing the SNP positions and the \(Z_{\beta} r^2/E[r^2]\) values for those SNPs

**References**

## Examples

```r
## load the snps and LDprofile example datasets
data(snps)
data(LDprofile)
## run Zbeta_rsq_over_expected over all the SNPs with a window size of 3000 bp
Zbeta_rsq_over_expected(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances,
                         LDprofile$bin,LDprofile$rsq)
## only return results for SNPs between locations 600 and 1500 bp
Zbeta_rsq_over_expected(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances,
                         LDprofile$bin,LDprofile$rsq,X=c(600,1500))
```

### Zbeta_Zscore

Runs the Zbeta function using the Z score of the r-squared values for the region

### Description

Returns a $Z_{\beta}^{\text{score}}$ value for each SNP location supplied to the function, based on the expected $r^2$ values given an LD profile and genetic distances. For more information about the $Z_{\beta}^{\text{score}}$ statistic, please see Jacobs (2016). The $Z_{\beta}^{\text{score}}$ statistic is defined as:

$$Z_{\beta}^{\text{score}} = \frac{\sum_{i \in L, j \in R} \frac{r_{i,j}^2 - E[r_{i,j}^2]}{\sigma[r_{i,j}^2]}}{|L||R|}$$

where $|L|$ and $|R|$ are the number of SNPs to the left and right of the current locus within the given window $ws$, $r^2$ is equal to the squared correlation between a pair of SNPs, $E[r^2]$ is equal to the expected squared correlation between a pair of SNPs, given an LD profile, and $\sigma[r^2]$ is the standard deviation.

### Usage

```r
Zbeta_Zscore(  
    pos,  
    ws,  
    x,  
    dist,  
    LDprofile_bins,  
    LDprofile_rsq,  
    LDprofile_sd,  
    minRandL = 4,  
    minRL = 25,  
    X = NULL  
)
```
Arguments

pos  A numeric vector of SNP locations

ws  The window size which the $Z_{\beta}^{\text{score}}$ statistic will be calculated over. This should be on the same scale as the pos vector.

x  A matrix of SNP values. Columns represent chromosomes; rows are SNP locations. Hence, the number of rows should equal the length of the pos vector. SNPs should all be biallelic.

dist  A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as pos.

LDprofile_bins  A numeric vector containing the lower bound of the bins used in the LD profile. These should be of equal size.

LDprofile_rsq  A numeric vector containing the expected $r^2$ values for the corresponding bin in the LD profile. Must be between 0 and 1.

LDprofile_sd  A numeric vector containing the standard deviation of the $r^2$ values for the corresponding bin in the LD profile.

minRandL  Minimum number of SNPs in each set R and L for the statistic to be calculated. Default is 4.

minRL  Minimum value for the product of the set sizes for R and L. Default is 25.

X  Optional. Specify a region of the chromosome to calculate $Z_{\beta}^{\text{score}}$ for in the format c(startposition, endposition). The start position and the end position should be within the extremes of the positions given in the pos vector. If not supplied, the function will calculate $Z_{\beta}^{\text{score}}$ for every SNP in the pos vector.

Details

The LD profile describes the expected correlation between SNPs at a given genetic distance, generated using simulations or real data. Care should be taken to utilise an LD profile that is representative of the population in question. The LD profile should consist of evenly sized bins of distances (for example 0.0001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

Value

A list containing the SNP positions and the $Z_{\beta}^{\text{score}}$ values for those SNPs

References


Examples

```r
## load the snps and LDprofile example datasets
data(snps)
data(LDprofile)
## run Zbeta_Zscore over all the SNPs with a window size of 3000 bp
Zbeta_Zscore(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances,
```
Zbeta_Zscore

LDprofile$bin,LDprofile$rsq,LDprofile$sd)
## only return results for SNPs between locations 600 and 1500 bp
Zbeta_Zscore(snps$positions,3000,as.matrix(snps[,3:12]),snps$distances,
LDprofile$bin,LDprofile$rsq,LDprofile$sd,X=c(600,1500))
Index

*Topic **datasets**
- LDprofile, 2
- snps, 4

L_plus_R, 3, 8
LDprofile, 2
LR, 2, 8
snps, 4

Zalpha, 5, 8
Zalpha_all, 6
Zalpha_BetaCDF, 8, 9
Zalpha_expected, 8, 11
Zalpha_log_rsq_over_expected, 8, 12
Zalpha_rsq_over_expected, 8, 14
Zalpha_Zscore, 8, 16
Zbeta, 8, 18
Zbeta_BetaCDF, 8, 19
Zbeta_expected, 8, 21
Zbeta_log_rsq_over_expected, 8, 22
Zbeta_rsq_over_expected, 8, 24
Zbeta_Zscore, 8, 26