Package 'zalpha'

January 15, 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>

Description A suite of statistics from the Zalpha family for identifying areas of the genome under selective pressure.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

Suggests testthat (>= 2.1.0)

R topics documented:

LR *Runs the LR function*

Description

Returns the |L||R| value for each SNP location supplied to the function. 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

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

L_plus_R *Runs the L_plus_R function*

Description

Returns the $\binom{|L|}{2} + \binom{|R|}{2}$ value for each SNP location supplied to the function. 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

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

`Zalpha`	*Runs the Zalpha function*

Description

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

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

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

Usage

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

Arguments

`pos`	A numeric vector of SNP locations
`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.
`ws`	The window size which the Z_α statistic will be calculated over. This should be on the same scale as the `pos` vector.
`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_α 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_α for every SNP in the `pos` vector.

Value

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

References

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

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, x = NULL, ws, 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
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.
ws	The window size which the statistics will be calculated over. This should be on the same scale as the pos vector.
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 Zalpha_expected, Zbeta_expected, LR and L_plus_R will be calculated.
For any of the statistics which use an expected r^2 value, the parameters dist, LDprofile_bins and LDprofile_rsq must be supplied. This includes the statistics: Zalpha_expected, Zalpha_rsq_over_expected, Zalpha_log_rsq_over_expected, Zalpha_Zscore, Zalpha_BetaCDF, Zbeta_expected, Zbeta_rsq_over_expected, Zbeta_log_rsq_over_expected, Zbeta_Zscore and Zbeta_BetaCDF.

- For Zalpha_Zscore and Zbeta_Zscore to be calculated, the parameter LDprofile_sd must also be supplied.
- For Zalpha_BetaCDF and Zbeta_BetaCDF to be calculated, the parameters LDprofile_Beta_a and 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

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

See Also

Zalpha Zalpha_expected Zalpha_rsq_over_expected Zalpha_log_rsq_over_expected Zalpha_Zscore Zalpha_BetaCDF Zbeta Zbeta_expected Zbeta_rsq_over_expected Zbeta_log_rsq_over_expected Zbeta_Zscore Zbeta_BetaCDF LR L_plus_R

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:

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

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

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

Arguments

pos   A numeric vector of SNP locations

x   A matrix of SNP values. Columns represent chromosomes, rows are SNP lo-
                cations. 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.

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.

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 corre-
                sponding bin in the LD profile.

LDprofile_Beta_b

                A numeric vector containing the second estimated Beta parameter for the corre-
                sponding 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, gener-
ated using simulations or real data. Care should be taken to utilise an LD profile which is represen-
tative of the population in question. The LD profile should consist of evenly-sized bins of distances
(for example 0.00001 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

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust
Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

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

Description

Returns a $Z_\alpha^{E[r^2]}$ 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 $Z_\alpha^{E[r^2]}$ statistic please see Jacobs (2016). The $Z_\alpha^{E[r^2]}$ statistic is defined as:

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

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

Usage

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

Arguments

pos	A numeric vector of SNP locations
dist	A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as pos.
ws	The window size which the $Z_\alpha^{E[r^2]}$ statistic will be calculated over. This should be on the same scale as the pos vector.
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 $Z_\alpha^{E[r^2]}$ 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 which is representative of the population in question. The LD profile should consist of evenly-sized bins of distances (for example 0.00001 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

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

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:

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

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, and $E[r^2]$ is equal to the expected squared correlation between a pair of SNPs, given an LD profile.

Usage

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

Arguments

pos	A numeric vector of SNP locations
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.
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.
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 $Z_{\alpha}^{log_{10}(r^2/E[r^2])}$ 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 $Z_{\alpha}^{log_{10}(r^2/E[r^2])}$ 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 which is representative of the population in question. The LD profile should consist of evenly-sized bins of distances (for example 0.00001 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

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

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:

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

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, and $E[r^2]$ is equal to the expected squared correlation between a pair of SNPs, given an LD profile.

Usage

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

Arguments

pos	A numeric vector of SNP locations
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.
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.
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 which is representative of the population in question. The LD profile should consist of evenly-sized bins of distances (for example 0.00001 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

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

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

Description

Returns a Z_{α}^{Zscore} 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_{α}^{Zscore} statistic please see Jacobs (2016). The Z_{α}^{Zscore} statistic is defined as:

$$Z_{\alpha}^{Zscore} = \frac{\binom{|L|}{2}^{-1} \sum_{i,j \in L} \frac{r_{i,j}^2 - E[r_{i,j}^2]}{\sigma[r_{i,j}^2]} + \binom{|R|}{2}^{-1} \sum_{i,j \in R} \frac{r_{i,j}^2 - E[r_{i,j}^2]}{\sigma[r_{i,j}^2]}}{2}$$

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

Arguments

pos	A numeric vector of SNP locations
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.
ws	The window size which the Z_{α}^{Zscore} statistic will be calculated over. This should be on the same scale as the pos vector.
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_{α}^{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_{α}^{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 which is representative of the population in question. The LD profile should consist of evenly-sized bins of distances (for example 0.00001 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_α^{Zscore} values for those SNPs

References

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

Zbeta   *Runs the Zbeta function*

Description

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

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

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

Usage

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

Arguments

pos	A numeric vector of SNP locations
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.
ws	The window size which the Z_β statistic will be calculated over. This should be on the same scale as the pos vector.
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_β 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_β for every SNP in the pos vector.

Value

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

References

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

Zbeta_BetaCDF	*Runs the Zbeta function using a cumulative beta distribution function on the r-squared values for the region*

Description

Returns a $Z_\beta^{BetaCDF}$ 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 $Z_\beta^{BetaCDF}$ statistic please see Jacobs (2016). The $Z_\beta^{BetaCDF}$ statistic is defined as:

$$Z_\beta^{BetaCDF} = \frac{\sum_{i \in L, j \in R} \frac{B(r_{i,j}^2; a, b)}{B(a,b)}}{|L||R|}$$

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 $\frac{B(r_{i,j}^2; a, b)}{B(a,b)}$ is the cumulative distribution function for the Beta distribution given the estimated a and b parameters from the LD profile.

Usage

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

Arguments

pos	A numeric vector of SNP locations
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.
ws	The window size which the $Z_\beta^{BetaCDF}$ statistic will be calculated over. This should be on the same scale as the pos vector.
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 $Z_\beta^{BetaCDF}$ 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.

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 which is representative of the population in question. The LD profile should consist of evenly-sized bins of distances (for example 0.00001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

Value

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

References

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

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

Description

Returns a $Z_\beta^{E[r^2]}$ 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 $Z_\beta^{E[r^2]}$ statistic please see Jacobs (2016). The $Z_\beta^{E[r^2]}$ statistic is defined as:

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

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

Usage

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

Arguments

pos	A numeric vector of SNP locations
dist	A numeric vector of genetic distances (e.g. cM, LDU). This should be the same length as pos.
ws	The window size which the $Z_\beta^{E[r^2]}$ statistic will be calculated over. This should be on the same scale as the pos vector.
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 $Z_\beta^{E[r^2]}$ 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 $Z_\beta^{E[r^2]}$ 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 which is representative of the population in question. The LD profile should consist of evenly-sized bins of distances (for example 0.00001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

Value

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

References

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

Zbeta_log_rsq_over_expected

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

---

Description

Returns a $Z_\beta^{log_{10}(r^2/E[r^2])}$ 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 $Z_\beta^{log_{10}(r^2/E[r^2])}$ statistic please see Jacobs (2016). The $Z_\beta^{log_{10}(r^2/E[r^2])}$ statistic is defined as:

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

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

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

Arguments

pos	A numeric vector of SNP locations
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.
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.
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(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 $Z_{\beta}^{log_{10}(r^2/E[r^2])}$ 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 which is representative of the population in question. The LD profile should consist of evenly-sized bins of distances (for example 0.00001 cM per bin), where the value given is the (inclusive) lower bound of the bin.

Value

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

References

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

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 $Z_\beta^{r^2/E[r^2]}$ 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 $Z_\beta^{r^2/E[r^2]}$ statistic please see Jacobs (2016). The $Z_\beta^{r^2/E[r^2]}$ statistic is defined as:

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

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

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

Arguments

pos	A numeric vector of SNP locations
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.
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.
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(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 $Z_\beta^{r^2/E[r^2]}$ 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 which is representative of the population in question. The LD profile should consist of evenly-sized bins of distances (for example 0.00001 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

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

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

Description

Returns a Z_β^{Zscore} 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_β^{Zscore} statistic please see Jacobs (2016). The Z_β^{Zscore} statistic is defined as:

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

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

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

Arguments

pos	A numeric vector of SNP locations
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.
ws	The window size which the Z_β^{Zscore} statistic will be calculated over. This should be on the same scale as the pos vector.
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_β^{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_β^{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 which is representative of the population in question. The LD profile should consist of evenly-sized bins of distances (for example 0.00001 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_β^{Zscore} values for those SNPs

References

Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807

Index