Package 'CRAVE'

January 17, 2013

Title CRAVE

Version 0.0.2

Date 2013-01-17

Author Joshua Sampson

Description An R package for detecting associations between rare variant SNPs and phenotypes in genome-wide association studies.

Maintainer William Wheeler <wheelerb@imsweb.com>

Depends R (>= 2.11.0)

License GPL-2

Archs i386, x64

R topics documented:

CRAVE 1
crave 2

Index **7**

CRAVE *CRaVe*

Description

An R package for detecting associations between rare variant SNPs and phenotypes in genome-wide association studies.

Details

The main function is `crave`. This function requires that the data be stored in two files. One file for the genotype data and one file for the covariate and outcome data. The genotype file must be either in TPED or VCF format. The phenotype file must be uncompressed. The package includes 13 different tests for identifying associations. In addition, user-defined tests can be written as R functions or C functions. The package also includes the hg19 data base of gene regions stored in the sampleData directory. This package includes the zlib library (http://zlib.net/) for reading genotype files compressed with gzip.

References

Ferguson J, Wheeler W, Fu Y, Prokunina L, Zhao H and Sampson J. Tests of Association for Rare Variants: A Common Framework

---

`crave` *crave*

Description

A function for detecting associations between rare variant SNPs and phenotypes

Usage

```
crave(geno_file, pheno_file, out_gene, out_snp, op=NULL)
```

Arguments

`geno_file`	The file containing the genotype data (either a TPED or VCF file). This file may be compressed with gzip.
`pheno_file`	The file containing the response and covariate data. If `geno_file` is a VCF file, then `pheno_file` must contain a column of subject ids.
`out_gene`	Output file for the gene results.
`out_snp`	Output file for the SNP results.
`op`	List of options. See `details`.

Details

Genotype data:
The first 4 columns of a TPED file should be the chromosome, SNP id, distance, and location. The distance in column 3 is not used, and the chromosome should be labeled as: 1-22, 23 or X, 24 or Y, 25 or XY, 26 or MT. Starting in column 5, are the genotypes for each subject. The order of these genotypes must match the order of the subjects in the phenotype file. An example row of a TPED file would be:
1 rs3132489 0 2345643 A A A G A A G G 0 0 A A A G
In the row above, 0 denotes a missing allele. Set the option `allele_miss` to specify a different missing value.

A VCF file must be grouped by chromosome and SNPs sorted by increasing location within each chromosome. The first 9 columns of a VCF file must be:
#CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT
Starting in column 10 are the subject ids. No order is assumed for these ids, however some of the ids must match the ids in the phenotype file. Subjects with un-matched ids will be removed from the anlaysis.

Phenotype data:
Any subject with a missing value for the response or a covariate will be removed from the analysis. The response can either be continuous or binary. If the response is binary and not coded as 0 for controls and 1 for cases, then use the options `control_value` and `case_value`.

Options list:
Below are the names for the options list `op`. All names have default values if they are not specified.

- adapt_maxperm Maximum number of adaptive permutations. Set to 0 for non-adaptive permutations. The stopping criteria (see adapt_tol) is checked after every nperm permutations. The default value is 1000000.
- adapt_tol Stopping tolerance for adaptive permutations.
  The algorithm stops if 3.92*sqrt(pval*(1.0-pval)/N)/pval < adapt_tol, where N is the total number of permutations that have been run so far. The default is 0.1.
- addcols 0-2. Specifying 1 or 2 will add columns to the out_gene file. The additional columns are the number of permutations and number of times that the permutation test statistic was greater than the observed. The default is 0.
- allele_miss Missing value for the alleles in a TPED file. The default value is 0.
- bm_n Number of points to consider in the Bounded Minimum Test. The default is 10.
- case_value Integer value of the response column denoting the cases. The default is 1.
- chr_alias Tab delimited text file with the first column as the chromosome number (1-26) and the second column the chromosome string. This option is only used with VCF genotype files when the chromosome fields are not labeled as integers.
- control_value Integer value of the response column denoting the controls. The default is 0.
- covars Vector of column names or column numbers in pheno_file to include as covariates. The default is that no covariates will be used.
- das_alpha A number between 0 and 1 for the alpha parameter in the Data Adaptive Sum Test. The default is 0.05.
- exclude_chr Vector of chromosome numbers to exclude in anlaysis. Example: 1-6,13,22. The default value is NULL.
- exclude_gene File containing 1 column of the genes to exclude in the analysis.
- exclude_snp File containing 1 column of the SNPs to exclude in the analysis.
- exclude_sub File containing 1 column of the subject ids to exclude in the analysis.
- gene_file Folder containing the gene information files. To use hg16, hg17, or hg18, set gene_file=folder containing the gene information files. By default, hg19 is used and is installed with CRaVe. The files are named chr_xxx.txt.gz. Each file contains the columns:
  Column 1: gene name
  Column 2: gene start position
  Column 3: gene end position
  Column 4: codon start position
  Column 5: codon end position
  Column 6: number of exon start and end positions
  Column 7: comma seperated list of exon starting positions
  Column 8: comma seperated list of exon ending positions
  The files must be tab-delimited, sorted by ascending gene start position.
- gene_kb Number of kilobases upstream and downstream to expand the gene regions in the gene tables. The default is 10.
- geno_gz 0 or 1 for an uncompressed/compressed genotype file with gzip. The default is determined from the file extension.
- geno_sep File delimiter for the genotype file.
- id Column name or number of the id variable in pheno_file. The default value is 1.
- include_chr Vector of chromosome numbers to include in anlaysis. Example: 1-6,13,22. The default value is 1:26

- include_gene File containing 1 column of the genes to include in the analysis.
- include_snp File containing 1 column of the SNPs to include in the analysis.
- include_sub File containing 1 column of the subject ids to include in the analysis.
- max_maf The maximum MAF for a SNP to be included in the analysis, so that any SNP with MAF > max_maf will be excluded. The default is 0.5.
- max_miss_rate Maximum missing rate for any SNP to be included in the analysis. The default value is 0.8.
- min_maf The minimum MAF for a SNP to be included in the analysis, so that any SNP with MAF < min_maf will be excluded. The default is 0.
- notGroupedByChr Set to TRUE if geno_file is a TPED file and this file is not grouped by chromosome. Note that a VCF file should always be grouped by chromosome. The default value is FALSE.
- no_gene_file Set to TRUE to not use the default gene table file for the gene names. Instead use the reflink.geneName value in the VCF file. This option is only valid if geno_file is a VCF file and the weight= option is not specified. The default is FALSE.
- nperm The number of permutations to run for each iteration of the adaptive permutation algorithm. The default is 500.
- pheno_header 0 or 1 if pheno_file contains column names.
- pheno_miss Missing value for the pheno_file. The default value is '.'
- pheno_sep File delimiter for the phenotype file.
- print The larger the value, the more information will be printed to the log_file. The default is 0 unless log_file is specified, in which case the default is 2.
- reflink.geneName Name of the variable in the INFO field of the VCF file that gives the gene name. The default value is "reflink.geneName"
- response Column name or number of the response variable in pheno_file. The default value is 2.
- rover_t Positive number t to define the weights (w = 1-exp(-t*delta*delta)) in the Rover Test. The default is 0.2
- seed Random seed. The default will be computed from the time.
- sortedByLoc 0 or 1. Set to 0 if the genotype data is not sorted by SNP location within each chromosome. The default is 1.
- store_perms 0 or 1. Set to 1 to store the permutations. Due to memory issues, this option should only be used for small sample sizes and number of permutations. The default is 0.
- tests Character vector of tests to use. The possible tests are:
  ALL - To compute all non user defined tests
  BM - Bounded Minimum Test
  CMC - Combined Multivariate and Collapsing Test
  DAS - Data Adaptive Sum Test
  FI - Fisher's Test
  HOI - Hotelling's T^2 Test assuming independence
  HOT - Hotelling's T^2 Test
  MDF - Similarity Regression Test
  MDP - Similarity Regression Test using positive deltas
  RO - ROVER (Reduction of Variables Explained by Random noise) Test
  RV - ROVER-V Rover Test that includes the variances in the weights
  STO - Stouffer's Z-score Test
  STP - Stouffer's Positive Z-score Test

SUM - Sum Test
TH - Threshold Test
SKAT - Alias for MDF [Under standard conditions, SKAT [PMID: 21737059] is equivalent to MDF]
CAL - Alias for C-alpha [Under standard conditions, the C-alpha statistic [PMID: 21408211] is equivalent to MDF]

Use the strings "c1", "c2", etc to specify user-defined C functions (see below). The default is to use the MDF test.

- `threshold_alpha` A number between 0 and 1 for the alpha parameter in the Threshold Test. The default is 0.05.
- `update_cor` Value such that the correlation matrix will be recomputed if the relative difference in the observed and permuted Hotelling test statistics is less than `update_cor`. A large value for `update_cor` will cause the correlation matrix and it's inverse to be recomputed in each permutation. A value less than or equal to 0 for `update_cor` will allow for quicker computation times. This option is only valid for case-control data. The default is 0.3.
- `user_func` Character vector of user-defined R function names to be used as test statistics. A maximum of 5 user-defined tests are allowed. See below for user-defined C functions.
- `vs` The value of `vs`, where `vs` abbreviates variance stabilization, is the value added to the estimated variance of each SNP. The default is 0.005
- `weight` "e", "c", or "n". weight=e assigns a weight of 1 to all exomic SNPs and 0 to other SNPs. weight=c assigns a weight of 1 to SNPs between the start and stop codon and 0 to other SNPs. weight=n assigns a weight of 1 to non-synonomous exomic SNPs and 0 to other SNPs. The default is that a weight of 1 is assigned to all SNPs. This option cannot be used with the `weight_file` option.
- `weight_file` File containing SNP weights. This file may have 1 or 2 columns only. For a 1 column file the column of weights must match the order of the SNPs in the genotype file. For a 2 column file, the first column is the SNP id and the second column the weight.
- `weight_header` Set to 1 if row 1 of the weight file contains column names. The default value is determined from the file.
- `weight_sep` One character value for the weight file delimiter.

User-defined tests:
Any user-defined R function to compute a test statistic must return a single number and have 4 input arguments. Argument 1 is the vector of scores for the SNPs in a gene, argument 2 is the gene covariance matrix, argument 3 is the gene correlation matrix, and argument 4 is the inverse correlation matrix for the gene. The results for these tests are in `out_file` with prefix "USER_R1", "USER_R2", etc, where the order corresponds to the order of the functions listed in the option `user_func`. A user-defined C function to compute a test statistic must be defined in the file `user_defined.c` in the "src" folder. Up to 5 user-defined C functions can be written, however the names of these functions must be "c1", "c2", "c3", "c4", "c5", and the order/number of input arguments cannot be modified. See the function "c1" in `user_defined.c`.

Value

The returned value is NULL. All final results are stored in `out_gene` and `out_snp`.

Examples

```
geno_file <- system.file("sampleData", "sim_tped.txt", package="CRAVE")
```

```
pheno_file <- system.file("sampleData", "sim_pheno.txt", package="CRAVE")
out_gene <- "c:/temp/out_gene.txt"
out_snp <- "c:/temp/out_snp.txt"

# Run with the defaults
#crave(geno_file, pheno_file, out_gene, out_snp, op=list(tests=c("RO", "SUM")))

# User defined R function
fnc1 <- function(delta, cov, cor, invcor) {
  return(sum(delta*delta))
}

# Include the user-defined test
#crave(geno_file, pheno_file, out_gene, out_snp, op=list(user_func="fnc1", print=3))
```

Index

∗Topic **model**
    crave, 2
∗Topic **package**
    CRAVE, 1

CRAVE, 1
crave, *1*, 2