Package 'JAGUAR'

March 3, 2015

Type Package

Title Joint Analysis of Genotype and Group-Specific Variability Using a Novel Score Test Approach to Map eQTL

Version 2.0

Date 2015-03-01

Author Chaitanya R. Acharya and Andrew S. Allen

Maintainer Chaitanya Acharya <c.acharya@duke.edu>

Depends R (>= 3.0.0), Rcpp, plyr, lme4, doParallel

LinkingTo Rcpp

NeedsCompilation yes

Description Implements a novel score test that measures 1) the overall shift in the gene expression due to genotype (additive genetic effect), and 2) group-specific changes in gene expression due to genotype (interaction effect) in a mixed-effects model framework.

License GPL-2

URL https://groups.google.com/d/forum/jaguar-r-package

Repository CRAN

Date/Publication 2015-03-03 00:58:02

R topics documented:

JAGUAR-package *Joint analysis of genotype and group-specific variability using a novel score test to map eQTL*

Description

The aim of the package is allow users to apply a novel score test method developed to map eQTL in the presence of multiple correlated groups (for example, tissues) from the same individual. We plan to do this by jointly analyzing all the groups by simultaneously measuring the total shift in the gene expression data due to genotypes and group-specific interaction of the genotypes with the gene expression data. Here is an example of a workflow.

1. We assume that the gene expression data and the genotype data are appropriately preprocessed. Usually, gene expression datasets are long and skinny, i.e. `p >> n`. We recommend to partition this gene expression data to run simultaneous analyses on all the partitions to save time. This can be performed using `SliceGeneData`

2. Run `jaguar` on each gene expression data partition to obtain a matrix of joint score test p-values with genes on rows and SNPs on columns.

3. A threshold p-value needs to be established to correct for multiplicity.

4. A threshold value can be computed from the above results using `calcThreshold` by combining results from all gene expression data analyses. We provide two methods to adjust for multiplicity. One is the standard Bonferroni correction and other's Danyu Lin's efficient Monte Carlo method. It is important to note that the results from all the partitions of the gene expression data are required to compute the threshold value by Lin's method.

5. This threshold value can be used to call the significant gene-SNP pairs in the analysis by running `ProcessJaguarResults` on the output from running `jaguar`.

6. Power or null simulations can be run using `jaguarSIM` by simulating one gene and one SNP at a time.

Details

Package:	JAGUAR
Type:	Package
Version:	2.0
Date:	2014-03-01
License:	GPL-2

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

An efficient Monte Carlo approach to assessing the statistical significance in genomic studies. Lin, D.Y. Bioinformatics. 21(6) 2005.

Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL. Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen (Manuscript submitted)

`calcThreshold` *Calculates various threshold values to control for the family-wise error rate (FWER).*

Description

Function to calculate thresholds values. Only Bonferroni and Danyu Lin's efficient Monte Carlo method are allowed.

Usage

```
calcThreshold(nsnps,ngenes,method=c("bonferroni","lin"),path,alpha=0.05)
```

Arguments

`nsnps`	An integer indicating the total number of SNPs in the study
`ngenes`	An integer indicating the total number of genes in the study
`path`	Path to all the directories containing mcMIN.txt files. This is essential when gene expression data is sliced into multiple fragments.
`method`	Choose method to adjust the FWER. Only Bonferroni and Lin's method are provided for now.
`alpha`	The nominal p-value indicating the false positive rate

Value

Numeric threshold value

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

An efficient Monte Carlo approach to assessing the statistical significance in genomic studies. Lin, D.Y. Bioinformatics. 21(6) 2005.

Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL. Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen (Manuscript submitted)

See Also

jaguar,ProcessJaguarResults,SliceGeneData,jaguar

Examples

```
## Example
#
# Load the example data (not an ideal dataset)
data(example.data);

# Set the parameter values
GeneExp = as.matrix(example.data$Gene);
GenoMat = as.matrix(example.data$Geno);
ngenes = nrow(GeneExp)
nsnps = nrow(GenoMat)

# Calculate the threshold values
calcThreshold(ngenes,nsnps,method="bonferroni",alpha=0.05);
```

---

example.data *An example gene expression and genotype data*

Description

This is a list object containing part simulated data set of 15 SNP values in allele dose format and 10 gene expression values for 10 patients in 4 groups.

Format

List containing gene expression data as a matrix with genes on rows and samples in columns, genotype data in allele dosage format with SNPs on rows and samples in columns, the number of groups in the study and a vector of sample ids.

Value

Gene	A matrix of gene expression data with 10 genes and 10 samples in four groups (so a total of 40 samples)
Geno	A matrix of genotype data with 15 SNPs and 10 samples with SNPs in allele dosage format i.e. 0, 1 or 2

ngroups	Number of groups in the study
Samples	A vector of sample ids in the study

---

gamma_test *An internal C++ function*

Description

Internal function that computes p-values from the variance component score test

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen; Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL (Manuscript submitted)

GENEapply *An internal C++ function*

Description

Internal function that computes the joint score test statistic over all the SNPs for all the genes

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen; Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL (Manuscript submitted)

getMinP *An internal C++ function*

Description

Internal function that computes the minimum vector over of the score test statistic p-values over all the genes using an efficient Monte Carlo approach.

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen; Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL (Manuscript submitted)

jaguar *Runs the joint score test statistic approach*

Description

Computes joint score test statistic to map group-specific expression quantitative trait loci (eQTL) that tests for the shifts in gene expression patterns due to genotype and variability among tissues in a mixed effects model framework.

Usage

```
jaguar(geneexp,geno,ngroups,write=FALSE)
```

Arguments

geneexp	A matrix of gene expression data with gene on rows and patient samples on columns. Missing values not allowed. There has to be equal number of samples in each group.
geno	A matrix of genotype data recoded as single allele dosage number (i.e. 0, 1 or 2) with rows representing SNPs and columns representing samples
ngroups	An integer representing the number of groups in the data
write	Boolean value indicating whether the results should be outputted into a tab-delimited text file

Value

A matrix of raw unadjusted p-values with rows representing genes and columns representing SNPs

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen; Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL (Manuscript submitted)

See Also

calcThreshold,lin,ProcessJaguarResults,SliceGeneData,jaguarSIM

Examples

```
# Load the example data
data(example.data);

# Set the parameters
GeneExp = as.matrix(example.data$Gene);
Geno = as.matrix(example.data$Geno);
k = example.data$ngroups;

# Run JAGUAR
out = jaguar(GeneExp,Geno,k,write=FALSE);
dim(out);
```

---

jaguarSIM *Run null or power simulations*

Description

Function to run power/null simulations by simulating one gene and one SNP at a time. The objective of these simulations is two pronged - 1) Check for the type I error control for the joint score test statistic, and 2) Compare two different null hypotheses where one's called a global null (bta=0 and PVEg =0) and other is local null (PVEg=0). Under the global null hypotheses, we fit a model where we assume that there is no main genotypic effect and group-specific variability in the data. Under the local null, we fit a model where we assume only the absence of group-specific variability. This leads to a 1 degree-of-freedom variance component score test.

Usage

```
jaguarSIM(nobs = 500, k = 5, tau = 1, eps = 1, PVEg = 0, bta = 0, maf = 0.10)
```

Arguments

nobs	The number of observations in each group
k	The total number of groups
tau	Variance component of the subject-specific random effect
eps	Variance component of the residual error
PVEg	Proportion of variance explained by gamma
bta	Additive genotypic effect as a fixed-effect
maf	Minor allele frequency

Details

This function currently implements only balanced designs with equal number of observations in each group. A linear mixed effects model as described in the manuscript is fit to run the analyses. Please refer to our manuscript for more details.

Value

A numeric vector consisting of two different p-values, "GammaScoreTest" and "JointScoreTest" with the former indicating the p-value from the variance component score test and the latter indicating the p-value from the joint score test.

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen; Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL (Manuscript submitted)

See Also

calcThreshold, lin, ProcessJaguarResults, SliceGeneData, jaguar

Examples

```
## An example to perform some null simulations
## NOTE: 10 sims are definitely not enough. Please try between 1000-10000.

nsim=10; alpha=0.05;
test = do.call("rbind",rlply(nsim,.progress="time",jaguarSIM(nobs=10,k=4)));
null.sim = apply(test,2,function(x) sum(x<=alpha)/nsim);
```

jag_param *An internal C++ function*

Description

Internal function that computes jaguar parameters.

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen; Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL (Manuscript submitted)

lin *Efficient Monte Carlo approach to adjust for multiplicity*

Description

Computes the weighted sum of independent standard normal random variables by multiplying the score test statistic with multiple realizations of normal random samples. Using an efficient Monte Carlo approach, we approximate the joint distribution of the novel joint score test statistic and then use the Monte Carlo distribution to evaluate the error rates of the statistic. This approach of multiple hypotheses correction in the case of genomic studies has been shown to be computationally inexpensive when compared with standard permutation based approaches.

Usage

```
lin(geneexp,geno,ngroups,mc.real=5000,parallel=FALSE,ncores=2,snp.slice=10000,write=TRUE)
```

Arguments

geneexp	A matrix of gene expression data with gene on rows and patient samples on columns. Missing values not allowed. There has to be equal number of samples in each group.
geno	A matrix of genotype data recoded as single allele dosage number (i.e. 0, 1 or 2) with rows representing SNPs and columns representing samples
ngroups	An integer representing the number of groups in the data
mc.real	An integer representing the number of Monte Carlo realizations of the score test statistic. Higher the number, higher the accuracy and longer the computation time
parallel	A boolean value indicating whether the analysis should be parallelized

ncores	An integer representing the number of cores used for parallel execution. The number must be at least 2. This option is required only when the analysis is parallelized
snp.slice	An integer representing the partition size of the genotype/SNP data. This number basically represents the number of SNPs analyzed at a time. This option is used only when the parallel execution of the analysis is activated.
write	A boolean value on whether the results should be written into a text file. By default, the results are stored in a text file 'mcMIN.txt' under each sub-directory

Value

A vector of length equal to the number of Monte Carlo realizations containing minimum p-values over all the realizations of the statistic

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

An efficient Monte Carlo approach to assessing the statistical significance in genomic studies. Lin, D.Y. Bioinformatics. 21(6) 2005.

Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL. Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen (Manuscript submitted)

See Also

calcThreshold, jaguar, ProcessJaguarResults, SliceGeneData, jaguarSIM

Examples

```
# Example
#
# Load the example data
data(example.data)

# Set the parameters
GeneExp = as.matrix(example.data$Gene);
Geno = as.matrix(example.data$Geno);
k = example.data$ngroups;

# Compute the minimum p-value over all the realizations of the score test statistic
# We recommend at least 5000 MC realizations

# This is just an  example
lin_mc = lin(GeneExp,Geno,k,mc.real=100,write=FALSE)
length(lin_mc)
```

plotqtl *Plot of eQTL results*

Description

Scatter plot displaying eQTL results with transcript location on the y-axis and SNP location on the x-axis. This plot is an implementation of ePlot function from Wei Sun's eMap R-package.

Usage

```
plotqtl(geneID,snpID,gene.chr,gene.pos,snp.chr,snp.pos,scores,chroms)
```

Arguments

geneID	A vector indicating the genes to be mapped
snpID	A vector indicating the SNPs to be mapped
gene.chr	A vector indicating the chromosomal location of the genes to be mapped
gene.pos	A vector indicating the start site of all the genes on the Gene Chip
snp.chr	A vector indicating the chromosomal location of the SNPs to be mapped
snp.pos	A vector indicating the chromosomal location of all the SNPs on the SNP Chip
scores	A vector of p-values of each Gene-SNP pair
chroms	A vector indicating the number of chromosomes to me mapped. Usually, it is 1 to 22 (excluding X and Y chromosomes)

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

eQTL analysis by Linear Model http://www.bios.unc.edu/~weisun/software/eMap.pdf

Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL. Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen (Manuscript submitted)

See Also

jaguar,ProcessJaguarResults,SliceGeneData,jaguar

Examples

```
## NOT RUN
### Read the annotation file of the Gene Chip
#genes = read.table("gene_annotation.txt",header=T,check.names=F)
#eChr = genes$Chromosome
#ePos = genes$StartSite
#
### Read the annotation file of the SNP Chip
#snps = read.table("snp_annotation.txt",header=F)
#mChr = snps$Chr
#mPos = snps$Pos
#
### Read the significant Gene-SNP pairs that are needed to be mapped
#out = ProcessJaguarResults(jaguar.out,threshold=0.05)
#
#geneID = match(out$Probes,genes$Probe_Id)
#markerID = match(out$SNPs,snps$SNP_Id)
#scores = out$P.value
#chroms=1:22
#
#plotqtl(geneID,snpID,gene.chr,gene.pos,snp.chr,snp.pos,scores,chroms)
```

`ProcessJaguarResults` *Obtain significant gene-SNP pairs based on a predetermined threshold value*

Description

Function that processes results from running jaguar and outputs gene-SNP pairs deemed significant by using a predetermined threshold value. It also has an option to print QQ-plot of the p-values from the analysis.

Usage

```
ProcessJaguarResults(jaguar.out,threshold,plot=FALSE)
```

Arguments

`jaguar.out`	A Matrix of joint score test statistic values with genes on rows and SNPs on columns
`threshold`	An integer representing a threshold value to call for significance
`plot`	Takes a Boolean value. If 'TRUE', prints a QQ-plot of the p-values from the analysis. In the interests of time and memory management, if there are more than 500,000 gene-SNP pairs in the analysis, only randomly selected 500,000 gene-SNP pairs will be plotted

Value

A matrix containing three columns – 1) Genes, 2) SNPs and 3) P-value from the joint score test approach

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL. Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen (Manuscript submitted)

See Also

calcThreshold,lin,jaguar,SliceGeneData,jaguarSIM

Examples

```
## Example
#
# Load the example data
data(example.data);
#
# Set the parameters
GeneExp = as.matrix(example.data$Gene);
Geno = as.matrix(example.data$Geno);
ngroups=example.data$ngroups;
#
# Run JAGUAR
jag.out = jaguar(GeneExp,Geno,ngroups,write=FALSE);
#
# Process JAGUAR output with a predetermined threshold value
# computed from ComputeLinThreshold()
#
result = ProcessJaguarResults(jag.out,0.50);
dim(result);
head(result);
```

RowMin *An internal C++ function*

Description

Internal function that computes row min of a matrix.

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen; Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL (Manuscript submitted)

`RowSums` *An internal C++ function*

Description

Internal function that computes row sums.

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen; Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL (Manuscript submitted)

`scoreTest` *An internal C++ function*

Description

Internal function that computes p-values from the joint score test.

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen; Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL (Manuscript submitted)

SliceGeneData *Slice gene expression data into multiple partitions*

Description

Function to 1) create sub-directories, 2) slice gene expression data into partitions of predetermined size, and 3) sliced gene expression partitions are deposited into each sub-directory

Usage

```
SliceGeneData(geneexp,size,path=getwd())
```

Arguments

geneexp	A matrix of gene expression data with gene on rows and patient samples on columns. Missing values not allowed. For now, there has to be an equal number of samples in each group.
size	Integer indicating the size of each slice of gene expression data.
path	Location for the sub-directories. Please give the full path. Default is set to the current directory.

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen; Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL (Manuscript submitted)

See Also

calcThreshold, lin, jaguar, ProcessJaguarResults, jaguarSIM

Examples

```
# Set the size of the partition
# size = 100; ## Indicates the number of genes in each partitioned gene exp data
#
# Assuming that the path is the default getwd()
# SliceGeneData(geneexp,size)
```

snpOUT *An internal C++ function*

Description

Internal function that computes the joint score test statistic over all the SNPs for a given gene

Author(s)

Chaitanya R. Acharya, Andrew S. Allen Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar and Andrew S. Allen; Joint analysis of genotype and tissue-specific variability using a novel score test to map eQTL (Manuscript submitted)

Index