Package 'JAGUAR'

August 29, 2016

Type Package

Title Joint Analysis of Genotype and Group-Specific Variability Using a Novel Score Test Approach to Map Expression Quantitative Trait Loci (eQTL)

Version 3.0.1

Date 2016-03-01

Author Chaitanya R. Acharya and Andrew S. Allen

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

Depends R (>= 3.0.0), Rcpp, plyr, lme4, reshape2

LinkingTo RcppArmadillo, Rcpp, RcppProgress

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 2016-07-27 02:52:37

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 `jaguar_slice`

2. If performing a genome-wide analysis, run `jaguar_gwa` on each gene expression data partition to obtain a matrix of joint score test p-values with genes on rows and SNPs on columns. If performing a cis analysis, run `jaguar_cis` on each gene expression data partition.

3. Permutation resampling can be performed while running cis analysis and gene-level p-values can be obtained. We do not recommend permutations for genome-wide analysis due to the computational burden.

4. After running a genome-wide analysis, `jaguar_process` function can be used to identify significant gene-SNP pairs based on a predetermined or user-defined threshold value.

5. Power or null simulations can be run using `jaguar_sim` by simulating one gene-SNP pair at a time.

Details

Package:	JAGUAR
Type:	Package
Version:	3.0.1
Date:	2016-07-11
License:	GPL-2

Author(s)

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

References

Chaitanya R. Acharya, Kouros Owzar, Janice M. McCarthy and Andrew S. Allen; Exploiting expression patterns across multiple tissues to map expression quantitative trait loci. BMC Bioinformatics (2016) 17:257 DOI 10.1186/s12859-016-1123-5

---

cis_eqtl *An internal C++ function*

Description

Internal function to perform cis-eQTL analysis

Author(s)

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

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

---

jagSIM *An internal C++ function*

Description

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

Author(s)

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

`jaguar_cis` *Perform cis-eQTL analysis*

Description

Computes p-value from our joint score test in a cis framework 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. A gene-level p-value is computed using a permutation-resampling scheme in order to investigate if a gene has at least one eQTL across all the groups.

Usage

```
jaguar_cis(geneexp,genomat,snp.bed,gene.bed,cisDist=100000,nperm=10000,seed=100)
```

Arguments

`geneexp`	A matrix of gene expression data with gene on rows and patient samples on columns. There has to be equal number of samples in each group. Samples (columns in the geneexp object) with missing gene expression values for any group/tissue MUST be included in the data
`genomat`	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
`snp.bed`	BED file format of SNP description. For more information, see http://genome.ucsc.edu/FAQ/FAQformat.html#format1 or the example data
`gene.bed`	BED file format of Gene description. For more information, see http://genome.ucsc.edu/FAQ/FAQformat.html#format1 or the example data
`cisDist`	cis distance is defined as the maximum absolute distance between the gene and a SNP. Default value is 100Kb
`nperm`	Number of permutations. Default value is 10,000. Note that if it is 0, no permutations are performed
`seed`	Seed value for permutations

Value

If permutations are performed, a list object with the following components is returned -

`results`	A matrix containing gene-level permutation p-values
`cisSNPs`	A list object containing the cis-SNPs corresponding to the genes in results

If permutations are not performed -

A matrix containing raw unadjusted p-values for all gene-cisSNP pairs

Author(s)

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

References

Chaitanya R. Acharya, Kouros Owzar, Janice M. McCarthy and Andrew S. Allen; Exploiting expression patterns across multiple tissues to map expression quantitative trait loci. BMC Bioinformatics (2016) 17:257 DOI 10.1186/s12859-016-1123-5

See Also

jaguar_process, jaguar_slice, jaguar_sim, jaguar_gwa, jaguar_plotqtl

Examples

```
# Load the example data
data(jaguar_example);
Gene = jaguar_example$GENE_EXP
SNP = jaguar_example$GENO_MAT
gene_loc = jaguar_example$GENE_BED
snp_loc = jaguar_example$SNP_BED
# Run a cis analysis with no permutations
out = jaguar_cis(Gene,SNP,snp_loc,gene_loc,nperm=0);
length(out)
```

---

jaguar_example *Simulated multi-tissue eQTL data*

Description

This is a list object containing a simulated eQTL data

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, gene and SNP information in BED format.

Value

GENE_EXP	A matrix of gene expression data with 100 genes and 100 individuals in five groups (so a total of 500 samples)
GENO_MAT	A matrix of genotype data with 1,036 SNPs in 100 individuals with SNPs in allele dosage format i.e. 0, 1 or 2
GENE_BED	Gene location information in BED file format
SNP_BED	SNP location information in SNP file format

jaguar_gwa *Perform genome-wide analysis*

Description

Computes p-value from our joint score test in a cis framework 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. A gene-level p-value is computed using a permutation-resampling scheme in order to investigate if a gene has at least one eQTL across all the groups.

Usage

```
jaguar_gwa(geneexp,genomat,write=FALSE)
```

Arguments

geneexp	A matrix of gene expression data with gene on rows and patient samples on columns. There has to be equal number of samples in each group. Samples (columns in the geneexp object) with missing gene expression values for any group/tissue MUST be included in the data
genomat	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
write	Boolean value indicating whether the results should be outputted into a tab-delimited text file. The default value is FALSE.

Value

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

Author(s)

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

References

Chaitanya R. Acharya, Kouros Owzar, Janice M. McCarthy and Andrew S. Allen; Exploiting expression patterns across multiple tissues to map expression quantitative trait loci. BMC Bioinformatics (2016) 17:257 DOI 10.1186/s12859-016-1123-5

See Also

jaguar_process, jaguar_slice, jaguar_sim, jaguar_cis, jaguar_plotqtl

Examples

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

# Run a cis analysis with no permutations
Gene_Mat = as.matrix(jaguar_example$GENE_EXP[1:10,])
Geno_Mat = as.matrix(jaguar_example$GENO_MAT)
out = jaguar_gwa(Gene_Mat,Geno_Mat);
dim(out)
```

`jaguar_plotqtl` *Plotting the 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

```
jaguar_plotqtl(geneID,snpID,gene.chr,gene.pos,snp.chr,snp.pos,scores,chroms=1:22)
```

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

References

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

Chaitanya R. Acharya, Kouros Owzar, Janice M. McCarthy and Andrew S. Allen; Exploiting expression patterns across multiple tissues to map expression quantitative trait loci. BMC Bioinformatics (2016) 17:257 DOI 10.1186/s12859-016-1123-5

See Also

jaguar_gwa, jaguar_process, jaguar_slice, jaguar_cis, jaguar_sim

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 = jaguar_process(jaguar.out,threshold=0.05)
#
#geneID = match(out$Gene,genes$Probe_Id)
#markerID = match(out$SNP,snps$SNP_Id)
#scores = out$P.value
#chroms=1:22
#
#jaguar_plotqtl(geneID,snpID,gene.chr,gene.pos,snp.chr,snp.pos,scores,chroms)
```

---

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

Description

Function that processes results from running a genome-wide analysis of 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

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

Arguments

jaguar.out	A Matrix of joint score test p-values with genes on rows and SNPs on columns
threshold	An numeric value 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 Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar, Janice M. McCarthy and Andrew S. Allen; Exploiting expression patterns across multiple tissues to map expression quantitative trait loci. BMC Bioinformatics (2016) 17:257 DOI 10.1186/s12859-016-1123-5

See Also

jaguar_gwa, jaguar_slice, jaguar_sim, jaguar_plotqtl, jaguar_cis

Examples

```
## Example
#
# Load the example data
data(jaguar_example);

# Genome-wide analysis
Gene_Mat = as.matrix(jaguar_example$GENE_EXP[1:10,])
Geno_Mat = as.matrix(jaguar_example$GENO_MAT)
jag.out = jaguar_gwa(Gene_Mat,Geno_Mat);
dim(jag.out);

# Process results based on a predetermined threshold
result = jaguar_process(jag.out,0.05);
dim(result);
```

---

jaguar_sim *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 is essentially a variance component score test.

Usage

```
jaguar_sim(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. For each individual, we model the potential genetic association between a target SNP and the expression level a target gene (in multiple tissues) at a single locus using the following mixed effects model (i = individual; t = tissue) -

$$y_{i,t} = \alpha_t + g_i\beta_i + b_t g_i + u_i + \epsilon_{i,t}$$

where $y_{i,t}$ is a t-dimensional vector of gene expression data for individual i, g_i is the scalar value of genotype in allele-dosage format, b_t is a t-dimensional tissue-specific random effect where $b \sim N(0, \gamma)$, u_i is the scalar value representing individual-specific random effect where $u \sim (0, \tau)$.

Value

A numeric vector consisting of two different p-values, "VCScoreTest" 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, Janice M. McCarthy and Andrew S. Allen; Exploiting expression patterns across multiple tissues to map expression quantitative trait loci (Manuscript submitted)

Chaitanya R. Acharya and Andrew S. Allen; JAGUAR: An R-package to Implement Joint Analysis of Genotype and Group-Specific Variability Using a Novel Score Test to Map eQTL (Manuscript submitted)

See Also

jaguar_process, jaguar_slice, jaguar_gwa, jaguar_cis, jaguar_plotqtl

Examples

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

nsim=10; alpha=0.05;
test = do.call("rbind",rlply(nsim,.progress="none",jaguar_sim(nobs=100,k=4)));
null.sim = apply(test,2,function(x) sum(x<=alpha)/nsim);
```

`jaguar_slice` *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

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

Arguments

`geneexp`	A matrix of gene expression data with gene on rows and patient samples on columns. There has to be equal number of samples in each group. Samples (columns in the geneexp object) with missing gene expression values for any group/tissue MUST be included in the data
`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 Maintainer: Chaitanya Acharya<c.acharya@duke.edu>

References

Chaitanya R. Acharya, Kouros Owzar, Janice M. McCarthy and Andrew S. Allen; Exploiting expression patterns across multiple tissues to map expression quantitative trait loci. BMC Bioinformatics (2016) 17:257 DOI 10.1186/s12859-016-1123-5

See Also

`jaguar_gwa`, `jaguar_process`, `jaguar_sim`, `jaguar_plotqtl`, `jaguar_cis`

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()
# jaguar_slice(geneexp,size)
```

`jag_fun` *An internal C++ function*

Description

Internal function that computes p-values from our joint score test

Author(s)

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

`rowsumscpp` *An internal C++ function*

Description

Internal function to compute row sums of a matrix.

Author(s)

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

`vcSIM` *An internal C++ function*

Description

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

Author(s)

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

Index