Package ‘PQLseq’

June 6, 2021

Type Package

Title Efficient Mixed Model Analysis of Count Data in Large-Scale Genomic Sequencing Studies

Version 1.2.1

Date 2021-06-06

Author Shiquan Sun, Jiaqiang Zhu, Xiang Zhou

Maintainer Jiaqiang Zhu <jiaqiang@umich.edu>

Description An efficient tool designed for differential analysis of large-scale RNA sequencing (RNAseq) data and Bisulfite sequencing (BSseq) data in the presence of individual relatedness and population structure. 'PQLseq' first fits a Generalized Linear Mixed Model (GLMM) with adjusted covariates, predictor of interest and random effects to account for population structure and individual relatedness, and then performs Wald tests for each gene in RNAseq or site in BSseq.

License GPL (>= 2)

Imports Rcpp (>= 0.12.14), foreach, doParallel, parallel, Matrix, methods

LinkingTo Rcpp, RcppArmadillo

NeedsCompilation yes

Depends R (>= 2.10)

Repository CRAN

Date/Publication 2021-06-06 13:30:02 UTC

R topics documented:

PQLseq-package .......................................................... 2
ExampleBSseq ............................................................. 3
ExampleRNAseq .......................................................... 4
pqlseq ................................................................. 4

Index 7
Description

An efficient tool designed for differential analysis of large-scale RNA sequencing (RNAseq) data and Bisulfite sequencing (BSseq) data in the presence of individual relatedness and population structure. 'PQLseq' first fits a Generalized Linear Mixed Model (GLMM) with adjusted covariates, predictor of interest and random effects to account for population structure and individual relatedness, and then performs Wald tests for each gene in RNAseq or site in BSseq. PQLseq is an R package for efficient differential analysis of large-scale RNA sequencing data and bisulfite sequencing data in the presence of individual relatedness and population structure. It first fits a Generalized linear mixed model with adjusted covariates, predictor of interest and random effects to account for population structure and individual relatedness, and then performs Wald test for each gene in RNA sequencing data or site in bisulfite sequencing data.

Details

<table>
<thead>
<tr>
<th>Package</th>
<th>PQLseq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type:</td>
<td>Package</td>
</tr>
<tr>
<td>Version:</td>
<td>1.10</td>
</tr>
<tr>
<td>Date:</td>
<td>2018-06-02</td>
</tr>
<tr>
<td>License:</td>
<td>GPL-3</td>
</tr>
</tbody>
</table>

Author(s)

Shiquan Sun, Jiaqiang Zhu, Xiang Zhou Maintainer: Shiquan Sun <shiquans@umich.edu>

References


ExampleBSseq

BSseq example dataset

Description
A simulated example dataset of BSseq for PQLseq.

Usage
data(ExampleBSseq)

Format
Contains the following objects:

- **mcount** a data frame containing the read counts for 5 sites.
- **predictor** a vector of 100 observations on a continuous variable.
- **relatednessmatrix** a genetic relationship matrix for 100 individuals.
- **totalcount** a data frame containing the total read counts for 5 sites.

Examples
data(ExampleBSseq)
attach(ExampleBSseq)
model_DNA=pqlseq(RawCountDataSet=mcount, Phenotypes=predictor, RelatednessMatrix=relatednessmatrix, LibSize=totalcount, fit.model="BMM",numCore=1)
head(model_DNA)
detach(ExampleBSseq)
ExampleRNAseq  
**RNAseq example dataset**

**Description**

A simulated example dataset of RNAseq for PQLseq.

**Usage**

data(ExampleRNAseq)

**Format**

Contains the following objects:

- `count` a data frame containing the read counts for 5 genes.
- `predictor` a vector of 100 observations on a continuous variable.
- `relatednessmatrix` a genetic relationship matrix for 100 individuals.
- `totalcount` a data frame containing the total read counts for 5 genes.

---

**pqlseq**  
*Fit Generalized Linear Mixed Model with Known Kinship Matrices Through Penalized-quasi Likelihood*

**Description**

Fit a generalized linear mixed model with a random intercept. The covariance matrix of the random intercept is proportional to a known kinship matrix.

**Usage**

```r
pqlseq(RawCountDataSet, Phenotypes, Covariates=NULL, RelatednessMatrix=NULL, LibSize=NULL, fit.model="PMM", fit.method = "AI.REML", fit.maxiter=500, fit.tol=1e-5, numCore=1, filtering=TRUE, verbose=FALSE,...)
```

**Arguments**

- `RawCountDataSet` a data frame containing the read count.
- `Phenotypes` a vector containing the predictor of interest.
- `Covariates` a data frame containing the covariates subject to adjustment (Default = NULL).
RelatednessMatrix

a known relationship matrix (e.g. kinship matrix in genetic studies). When supplied with a matrix, this matrix should be a positive semi-definite matrix with dimensions equal to the sample size in count data, and the order of subjects in this matrix should also match the order of subjects in count data. Currently there is no ID checking feature implemented, and it is the user’s responsibility to match the orders.

LibSize

a data frame containing the total read count. For Poisson mixed model, it will be calculated automatically if users do not provide. For binomial mixed model, it is required.

fit.model

describes the error distribution and link function to be used in the model. Either "PMM" for Poisson model, or "BMM" for binomial model (default = "PMM").

fit.method

method of fitting the generalized linear mixed model, currently only "REML" version is available.

fit.maxiter

a positive integer specifying the maximum number of iterations when fitting the generalized linear mixed model (default = 500).

fit.tol

a positive number specifying tolerance, the difference threshold for parameter estimates below which iterations should be stopped (default = 1e-5).

numCore

a positive integer specifying the number of cores for parallel computing (default = 1).

filtering

da logical switch for RNAseq data. By default, for each gene, at least two individuals should have read counts greater than 5. Otherwise, the gene is filtered (default = TRUE).

verbose

a logical switch for printing detailed information (parameter estimates in each iteration) for testing and debugging purpose (default = FALSE).

... additional arguments that could be passed to glm.

Details

Generalized linear mixed models (GLMM) are fitted using the penalized quasi-likelihood (PQL) method proposed by Breslow and Clayton (1993). Statistical inference in GLMM is notoriously difficult because of an intractable high-dimensional integral in the likelihood (Chen, 2016 and Lea, 2015), and by default we use the Average Information REML algorithm (Gilmour, Thompson and Cullis, 1995; Yang et al., 2011) to fit the model. An eigen-decomposition is performed in each outer iteration and the estimate of the variance component parameter $\tau$ is obtained by maximizing the profiled log restricted likelihood. When the Average Information REML algorithm fails to converge, a warning message is given and the algorithm is default to INLA approaches (Rue, 2009).

Value

numIDV number of individuals with data being analyzed
beta the fixed effect parameter estimate for the predictor of interest.
se_beta the standard deviation of fixed effect.
pvalue P value for the fixed effect, based on the Wald test.
h2    heritability of the transformed rate.
sigma2 total variance component.
overdisp dispersion parameter estimate
converged a logical indicator for convergence.

Author(s)
Shiquan Sun, Jiaqiang Zhu, Xiang Zhou

References

Examples
```r
data(ExampleRNAseq)
attach(ExampleRNAseq)
model_RNA=pqlseq(RawCountDataSet=count, Phenotypes=predictor,
                 RelatednessMatrix=relatednessmatrix, LibSize=totalcount,
                 fit.model="PMM",numCore=1)
head(model_RNA)
detach(ExampleRNAseq)
```
Index

* GLMMs
  pqlseq, 4
* datasets
  ExampleBSseq, 3
  ExampleRNAseq, 4
* function
  pqlseq, 4
* package
  PQLseq-package, 2
* pqlseq
  pqlseq, 4

ExampleBSseq, 3
ExampleRNAseq, 4

PQLseq (PQLseq-package), 2
pqlseq, 4
PQLseq-package, 2