Package 'CSeQTL'

July 21, 2025

Type Package

Title Cell Type-Specific Expression Quantitative Trail Loci Mapping

Version 1.0.0

Date 2025-03-04

Description Perform bulk and cell type-specific expression quantitative trail loci mapping with our novel method (Little et al. (2023) <doi:10.1038/s41467-023-38795-w>).

Encoding UTF-8

Imports Rcpp, smarter, ggplot2, multcomp, emdbook, MatrixEQTL, data.table, HelpersMG, R.utils, GenomicFeatures, methods

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 7.2.3

License GPL (>= 3)

Suggests rmarkdown, knitr

VignetteBuilder knitr

NeedsCompilation yes

Author Paul Little [aut, cre]

Maintainer Paul Little <pllittle321@gmail.com>

Depends R (>= 3.5.0)

Repository CRAN

Date/Publication 2025-03-06 17:00:06 UTC

Contents

CSeQTL_dataGen *CSeQTL_dataGen*

Description

Simulates a gene/SNP pair with baseline covariates `XX`, cell type compositions `true_RHO`, phased SNP genotypes `true_SNP`, and total (TReC) and allele-specific read counts (ASReC) contained in `dat`.

Usage

```
CSeQTL_dataGen(
  NN,
  MAF,
  true_BETA0 = log(1000),
  true_KAPPA,
  true_ETA,
  true_PHI = 0.1,
  true_PSI = 0.05,
  prob_phased = 0.05,
  true_ALPHA = NULL,
  batch = 1,
  RHO = NULL,
  cnfSNP = FALSE,
  show = TRUE
)
```

Arguments

`NN`	Positive integer for sample size.
`MAF`	Positive numeric value between 0 and 1 for the minor allele frequency to simulate phased SNP genotypes assuming Hardy-Weinberg.
`true_BETA0`	A positive numeric value denoting the reference cell type and reference base's expression multiplied by two and log transformed. For example, if the TReC for reference base and cell type is 500, then `true_BETA0` = `log(2 * 500)`.
`true_KAPPA`	A numeric vector denoting the baseline fold change in TReC between a cell type and reference. By definition, the first element is 1.
`true_ETA`	A numeric vector where each element denotes the fold change in TReC between the non-reference and reference base in a cell type.

true_PHI	A non-negative numeric value denoting the over-dispersion term associated with TReC. If true_PHI > 0, TReC is simulated with the negative binomial. If true_PHI = 0, TReC is simulated with the poisson.
true_PSI	A non-negative numeric value denoting the over-dispersion term associated with ASReC. If true_PSI > 0, ASReC is simulated with the beta-binomial, otherwise it is simulated with the binomial distribution.
prob_phased	A positive numeric value denoting the simulated proportion of simulated TReC that are ASReC.
true_ALPHA	By default, it is set to NULL setting each cell type with an eQTL to be cis-eQTL. Otherwise, a positive numeric vector of fold changes between TReC eQTL effect sizes and ASReC eQTL effect sizes.
batch	A numeric value set to 1 by default to allow underlying batch effects. Set to zero to eliminate batch effects.
RHO	A numeric matrix of cell type proportions where each row sums to one. If set to NULL, a matrix of cell type proportions will be simulated.
cnfSNP	A boolean value where TRUE re-arranges simulated SNPs to correlate with baseline bulk expression. When fitting the marginal model (not accounting for cell type proportions) and in the presence of cell type-specific differentiated expression, a marginal eQTL may be incorrectly inferred.
show	A boolean value to display verbose output and plot intermediate simulated results.

Value

A R list containing true parameters governing the simulated dataset, simulated covariate matrix XX, observed outcomes in dat.

CSeQTL_full_analysis *CSeQTL_full_analysis*

Description

Performs marginal and cell type-specific eQTL analysis with both CSeQTL and OLS models for simulation and comparative purposes.

Usage

```
CSeQTL_full_analysis(
  TREC,
  hap2,
  ASREC,
  PHASE,
  SNP,
  RHO,
  XX,
```

```
  log_lib_size,
  vec_MARG = c(TRUE, FALSE),
  vec_TRIM = c(TRUE, FALSE),
  vec_PERM = c(TRUE, FALSE),
  thres_TRIM = 10,
  ncores = 1,
  show = TRUE
)
```

Arguments

TREC	An integer vector of total read counts.
hap2	An integer vector of second haplotype counts
ASREC	An integer vector of total haplotype counts
PHASE	An integer vector of 0s and 1s denoting if a subject has available haplotype counts.
SNP	An integer vector of phased genotypes coded 0 (AA), 1 (AB), 2 (BA), 3 (BB), and 5 (NA).
RHO	A numeric matrix of cell type proportions. Rows correspond to subjects and columns correspond to cell types.
XX	A numeric design matrix of baseline covariates including the intercept in the first column and centered continuous covariates.
log_lib_size	A positive numeric vector of log transformed library sizes per subject.
vec_MARG	A boolean vector for marginal and/or cell type-specific analyses to be run. By default, both sets of analyses are run.
vec_TRIM	A boolean vector for whether or not analyses with trimmed outcomes are included. By default, both sets of analyses are run.
vec_PERM	A boolean vector for whether or not permuted SNP analyses are included. By default, both sets of analyses are run.
thres_TRIM	A positive numeric value to perform subject outcome trimming. Subjects with standardized Cooks' Distances greater than the threshold are trimmed.
ncores	A positive integer specifying the number of threads available to decrease computational runtime.
show	A boolean value to display verbose output and plot intermediate simulated results.

Value

A R list containing multiple objects. `res` is a R dataframe containing the model fitted, marginal model indicator, TRIM indicator, permutation indicator, cell types, utilized allele-specific reads indicator, A and B allele-specific expression, eta/eqtl fold change estimate, p-value. `out` contains lists of detailed estimates per model fit.

CSeQTL_GS *CSeQTL_GS*

Description

Main function that performs eQTL mapping on one gene with its associated SNPs.

Usage

```
CSeQTL_GS(
  XX,
  TREC,
  SNP,
  hap2,
  ASREC,
  PHASE,
  RHO,
  trim = TRUE,
  thres_TRIM = 20,
  numAS = 5,
  numASn = 5,
  numAS_het = 5,
  cistrans = 0.01,
  ncores = 1,
  show = TRUE
)
```

Arguments

XX	A numeric design matrix of baseline covariates including the intercept in the first column and centered continuous covariates. One of the non-intercept columns should correspond to centered log-transformed library size. The `rownames(XX)` needs to be specified.
TREC	An integer vector containing one gene's total read counts. The `names(TREC)` needs to be specified.
SNP	A nonnegative integer matrix with genotypes. Values should be coded 0, 1, 2, 3, 5 for genotypes AA, AB, BA, BB, NA, respectively. Rows correspond to SNPs and columns correspond to subjects. The `rownames(SNP)` needs to be specified.
hap2	An integer vector of the second haplotype's counts for the gene. The `names(hap2)` needs to be specified and match `names(TREC)`.
ASREC	An integer vector of the total haplotype counts (ASReC) for the gene. The `names(ASREC)` needs to be specified and match `names(TREC)`.
PHASE	A binary 0/1 vector indicating if haplotype counts for the gene will be used. The `names(PHASE)` needs to be specified and match `names(TREC)`.

RHO	A numeric matrix of cell type proportions. Rows correspond to subjects and columns correspond to cell types. The rownames(RHO) needs to be specified and match names(TREC). The colnames(RHO) also needs to be specified.
trim	Boolean value set to FALSE by default to prevent outcome trimming. If TRUE, the CSeQTL model will be fitted without SNP genotype to calculate each subject's Cooks' distance for the gene.
thres_TRIM	A positive numeric value to perform subject outcome trimming. Subjects with standardized Cooks' Distances greater than the threshold are trimmed.
numAS	A positive integer to determine if a subject has enough total haplotype counts.
numASn	A positive integer to determine how many subjects have at least numAS to use the haplotype counts.
numAS_het	A positive integer to determine how many subjects with at least numAS are heterozygous (AB or BA). If sum(PHASE == 1 & ASREC >= numAS & (SNP == 1 | SNP == 2)) >= numAS_het, those subjects haplotype counts will be used for TReCASE and cis/trans testing and estimation.
cistrans	A numeric value specifying the cis/trans test p-value cutoff to determine if the eQTLs from TReC-only or TReCASE model is reported.
ncores	A positive integer specifying the number of threads available to decrease computational runtime when performing trimming and looping through SNPs.
show	A boolean value to display verbose output.

Value

A R list of statistics and metrics after optimizing the model across multiple SNPs. The list contains parameter MLEs, gene expression estimates, fold change estimates, convergence indicators, likelihood ratio test statistics, and associated p-values.

CSeQTL_linearTest *CSeQTL_linearTest*

Description

Runs marginal and cell type-specific analysis using ordinary least squares.

Usage

```
CSeQTL_linearTest(
  input,
  XX,
  RHO,
  SNP,
  YY = NULL,
  MARG = FALSE,
  impute_geno = FALSE,
  trim = FALSE,
```

```
  thres_TRIM = 10,
  show_plot = TRUE,
  main_plot = "",
  CTs = NULL
)
```

Arguments

input	A data.frame containing columns total (total read counts) and log_lib_size (log transformed library size).
XX	A numeric design matrix of baseline covariates including the intercept in the first column and centered continuous covariates.
RHO	A numeric matrix of cell type proportions. Rows correspond to subjects and columns correspond to cell types.
SNP	An integer vector of phased genotypes coded 0 (AA), 1 (AB), 2 (BA), 3 (BB), and 5 (NA).
YY	Default is NULL. By default, the outcome for OLS is the inverse rank quantile normalized TReC after library size correction. Otherwise, the user can input their own transformed outcome as a numeric vector.
MARG	Boolean value. Set to TRUE to fit the marginal OLS model. Default is set to FALSE to fit the cell type-specific interaction OLS model.
impute_geno	Boolean value. Default is set to FALSE to only analyze subjects with non-missing genotype. If TRUE, missing genotypes are imputed with the mean of non-missing genotypes mimicing MatrixEQTL.
trim	Boolean value set to FALSE by default to prevent outcome trimming. If TRUE, the OLS model will be fitted without covariates containing SNP genotype to calculate each subject's Cooks' distance.
thres_TRIM	A positive numeric value to perform subject outcome trimming. Subjects with standardized Cooks' Distances greater than the threshold are trimmed.
show_plot	Boolean value set to TRUE to visualize boxplot or interactions.
main_plot	Character string for the visual's main title.
CTs	Set to NULL by default. If NULL and show_plot = TRUE, all cell type interaction plots are shown. Otherwise a subset of cell types can be displayed using an integer vector of columns or character vector of column names of RHO can be provided.

Value

A R list containing lm() output for lm_out, a R dataframe for out_df containing regression estimates, standard errors, p-values. res_trim provides a summary of trimmed results over a grid of cut-off values and number of samples trimmed. cooksd is a numeric vector of median shifted and MAD scaled Cook's distances per sample. prop_trim, a numeric value for number of samples with outcome values trimmed for the user-specified thres_TRIM value.

CSeQTL_oneExtremeSim *CSeQTL_oneExtremeSim*

Description

Performs a simulation with multiple replicates for a pre-specified set of arguments.

Usage

```
CSeQTL_oneExtremeSim(
  NN,
  MAF,
  true_BETA0,
  true_KAPPA,
  true_ETA,
  true_PHI = 0.1,
  wRHO,
  noiseRHO = 0,
  RR = 50,
  vec_MARG = c(TRUE, FALSE),
  vec_TRIM = c(TRUE, FALSE),
  vec_PERM = c(TRUE, FALSE),
  thres_TRIM = 10,
  ncores = 1,
  sim_fn
)
```

Arguments

NN	Positive integer for sample size.
MAF	Positive numeric value between 0 and 1 for the minor allele frequency to simulate phased SNP genotypes assuming Hardy-Weinberg.
true_BETA0	A positive numeric value denoting the reference cell type and reference base's expression multiplied by two and log transformed. For example, if the TReC for reference base and cell type is 500, then true_BETA0 = log{2 * 500}.
true_KAPPA	A numeric vector denoting the baseline fold change in TReC between a cell type and reference. By definition, the first element is 1.
true_ETA	A numeric vector where each element denotes the fold change in TReC between the non-reference and reference base in a cell type.
true_PHI	A non-negative numeric value denoting the over-dispersion term associated with TReC. If true_PHI > 0, TReC is simulated with the negative binomial. If true_PHI = 0, TReC is simulated with the poisson.
wRHO	Takes integer values 1, 2, or 3 to simulate three scenarios of cell type proportions.
noiseRHO	A positive numeric value to purposely distort simulated cell type compositions.

RR	A positive integer for number of replicates to generate and analyze.
vec_MARG	A boolean vector for marginal and/or cell type-specific analyses to be run. By default, both sets of analyses are run.
vec_TRIM	A boolean vector for whether or not analyses with trimmed outcomes are included. By default, both sets of analyses are run.
vec_PERM	A boolean vector for whether or not permuted SNP analyses are included. By default, both sets of analyses are run.
thres_TRIM	A positive numeric value to perform subject outcome trimming. Subjects with standardized Cooks' Distances greater than the threshold are trimmed.
ncores	A positive integer specifying the number of threads available to decrease computational runtime.
sim_fn	Character value specifying the full path and filename to store intermediate simulation replicates should errors arise.

Value

A R list containing the results of a small-scale simulation. res contains replicate-specific results while ures contains overall simulation results include power, frequency of replicates with constrained eta estimates and gene expressions inferred to be zero after optimization.

CSeQTL_run_MatrixEQTL *CSeQTL_run_MatrixEQTL*

Description

CSeQTL_run_MatrixEQTL

Usage

```
CSeQTL_run_MatrixEQTL(TREC, RD, XX, SNP, out_cis_fn, cisDist = 1e+06)
```

Arguments

TREC	A matrix of integer total read counts, rows are genes with row labels with syntax "gene_name:chrom:start:end", columns are samples with column labels.
RD	A positive numeric vector of library sizes per sample.
XX	A numeric design matrix of baseline covariates including the intercept in the first column and centered continuous covariates.
SNP	An integer matrix, rows are genomic loci with row labels such as "chrom:pos", columns correspond to samples with column labels.
out_cis_fn	A full path filename string to store MatrixEQTL output
cisDist	A positive integer for number of SNPs to include relative to the gene body

Value

The MatrixEQTL output file/R dataframe generated containing genes, SNPs, associated p-values, effect sizes, etc.

CSeQTL_smart *CSeQTL_smart*

Description

A function to understand the novelties within CSeQTL. The user can experiment with trimming TReC, assess parameter estimation, perform hypothesis testing, actively constrain cell type-specific parameters, when analyzing a single gene/SNP pair.

Usage

```
CSeQTL_smart(
  TREC,
  hap2,
  ASREC,
  PHASE,
  SNP,
  RHO,
  XX,
  upPHI,
  upKAPPA,
  upETA,
  upPSI,
  upALPHA,
  iFullModel = FALSE,
  trim = FALSE,
  thres_TRIM = 10,
  hypotest = TRUE,
  swap = TRUE,
  numAS = 5,
  numASn = 5,
  numAS_het = 5,
  cistrans_thres = 0.01,
  gr_eps = 0.01,
  conv_eps = 0.001,
  ncores = 1,
  show = FALSE
)
```

Arguments

TREC	An integer vector of total read counts.
hap2	An integer vector of second haplotype counts
ASREC	An integer vector of total haplotype counts
PHASE	An integer vector of 0s and 1s denoting if a subject has available haplotype counts.

SNP	An integer vector of phased genotypes coded 0 (AA), 1 (AB), 2 (BA), 3 (BB), and 5 (NA).
RHO	A numeric matrix of cell type proportions. Rows correspond to subjects and columns correspond to cell types.
XX	A numeric design matrix of baseline covariates including the intercept in the first column and centered continuous covariates.
upPHI	A value of 0 or 1 indicating if a Poisson or Negative binomial distribution is fitted, respectively.
upKAPPA	An integer vector of zeroes and ones where `length(upKAPPA) == ncol(RHO)`. A requirement is `upKAPPA[1] = 1`. Cell types with indices equal to one have the baseline fold change between the q-th and reference cell type are estimated. Otherwise that cell type's parameter is constrained to zero.
upETA	An integer vector of zeroes and ones where `length(upETA) == ncol(RHO)` indicating which cell types eQTL parameters are estimated or constrained to their null.
upPSI	A value of 0 or 1 indicating if a Binomial or Beta-binomial distribution is fitted, respectively.
upALPHA	An integer vector of zeroes and ones where `length(upALPHA) == ncol(RHO)` indicating which cell types' cis-trans parameters are estimated or constrained to their null.
iFullModel	A boolean that if set to TRUE will determine the submodel of the full model, based by upKAPPA, upETA, and upALPHA parameters, that can be estimated with stability.
trim	Boolean value. If TRUE, the TReC model will be fitted without SNP genotype to calculate each subject's Cooks' distance.
thres_TRIM	A positive numeric value to perform subject outcome trimming. Subjects with standardized Cooks' Distances greater than the threshold are trimmed.
hypotest	A boolean to perform eQTL significance testing and cis-trans eQTL testing.
swap	A boolean to determine if the reference cell type should be swapped with the cell type with highest TReC across alleles.
numAS	A positive integer to determine if a subject has enough total haplotype counts.
numASn	A positive integer to determine how many subjects have at least numAS to use the haplotype counts.
numAS_het	A positive integer to determine how many subjects with at least numAS are heterozygous (AB or BA). If `sum(PHASE == 1 & ASREC >= numAS & (SNP == 1 | SNP == 2)) >= numAS_het`, those subjects haplotype counts will be used for TReCASE and cis/trans testing and estimation.
cistrans_thres	A numeric value to determine the cis/trans test p-value cutoff.
gr_eps	A numeric value to determine if convergence is achieved based on the L2 norm of the gradient.
conv_eps	A numeric value to determine if convergence is achieved based on the L2 norm of the product of the inverse hessian and gradient.

ncores	A positive integer specifying the number of threads available to decrease computational runtime.
show	A boolean value to display verbose output and plot intermediate simulated results.

Value

A R list of statistics and metrics after optimizing the model for a single SNP. The list contains parameter MLEs, gradient vectors, covariance matrices, gene expression estimates, fold change estimates, convergence indicators, likelihood ratio test statistics, and associated p-values.

`gen_true_RHO` *gen_true_RHO*

Description

Simulates cell type proportions under three scenarios.

Usage

```
gen_true_RHO(wRHO = 1, NN, QQ, RHO = NULL)
```

Arguments

wRHO	An integer taking values 1, 2, or 3 for one of the three scenarios. For wRHO equal to 2 or 3, QQ needs to be set to 3 cell types.
NN	Positive integer for sample size.
QQ	Positive integer for number of cell types.
RHO	Default to NULL leads to simulating cell type proportions. If a matrix of proportions is supplied, this function will append row/column names.

Value

A numeric matrix with QQ columns and NN rows of cell type proportions per sample and cell type.

OLS_sim *OLS_sim*

Description

OLS_sim

Usage

```
OLS_sim(
  NN,
  MAF,
  true_BETA0,
  true_KAPPA,
  true_ETA,
  wRHO,
  RR,
  trim = FALSE,
  thres_TRIM = 10,
  showRHO = TRUE
)
```

Arguments

NN	Positive integer for sample size.
MAF	Positive numeric value between 0 and 1 for the minor allele frequency to simulate phased SNP genotypes assuming Hardy-Weinberg.
true_BETA0	A positive numeric value denoting the reference cell type and reference base's expression multiplied by two and log transformed. For example, if the TReC for reference base and cell type is 500, then `true_BETA0 = log{2 * 500}`.
true_KAPPA	A numeric vector denoting the baseline fold change in TReC between a cell type and reference. By definition, the first element is 1.
true_ETA	A numeric vector where each element denotes the fold change in TReC between the non-reference and reference base in a cell type.
wRHO	Takes integer values 1, 2, or 3 to simulate three scenarios of cell type proportions.
RR	A positive integer for number of replicates to generate and analyze.
trim	Boolean value set to FALSE by default to prevent outcome trimming. If TRUE, the OLS model will be fitted without covariates containing SNP genotype to calculate each subject's Cooks' distance.
thres_TRIM	A positive numeric value to perform subject outcome trimming. Subjects with standardized Cooks' Distances greater than the threshold are trimmed.
showRHO	Boolean for seeing a preview of how proportions are distributed.

Value

Simulation results for OLS-based approach with per replicate and per cell type metrics including p-values, OLS effect sizes, and proportion of samples with trimmed outcomes.

`plot_RHO` *plot_RHO*

Description

plot_RHO

Usage

```
plot_RHO(RHO, main_plot = "", ...)
```

Arguments

`RHO`	A numeric matrix of cell type proportions. Rows correspond to subjects and columns correspond to cell types.
`main_plot`	Character string for the visual's main title.
`...`	Arguments for `plot(x,y,bty = "n",...)`.

Value

Null. A plot of cell type proportions per cell type across `NN` samples.

`prep_gene_info` *prep_gene_info*

Description

Prepare GTF files for gene and exon level information

Usage

```
prep_gene_info(work_dir, gtf_gz_fn = NULL)
```

Arguments

`work_dir`	A character string specifying the working directory to store gene and exon data.
`gtf_gz_fn`	A character string full path of the gz compressed gtf file

Value

Null. No value returned.

Index