Package ‘xseq’

September 11, 2015

Title Assessing Functional Impact on Gene Expression of Mutations in Cancer

Version 0.2.1

Date 2015-08-25

Author Jiarui Ding, Sohrab Shah

Maintainer Jiarui Ding <jiaruid@cs.ubc.ca>

Depends R (>= 3.1.0)

Imports e1071 (>= 1.6-4), gptk (>= 1.08), impute (>= 1.38.1), preprocessCore (>= 1.26.1), RColorBrewer (>= 1.1-2), sfsmisc (>= 1.0-27)

Suggests knitr

Description
A hierarchical Bayesian approach to assess functional impact of mutations on gene expression in cancer. Given a patient-gene matrix encoding the presence/absence of a mutation, a patient-gene expression matrix encoding continuous value expression data, and a graph structure encoding whether two genes are known to be functionally related, xseq outputs: a) the probability that a recurrently mutated gene g influences gene expression across the population of patients; and b) the probability that an individual mutation in gene g in an individual patient m influences expression within that patient.

License GPL (>= 2)

LazyLoad true

VignetteBuilder knitr

LazyData TRUE

NeedsCompilation yes

Repository CRAN

Date/Publication 2015-09-11 08:04:31

R topics documented:
cna.call .......................................................... 2
cna.logr .......................................................... 3
Description

A dataset containing part of the The Cancer Genome Atlas acute myeloid leukemia Affymetrix SNP6.0 array copy number alteration calls from GISTIC

Usage

cna.call

Format

A matrix containing the GISTIC copy number calls of 454 genes in 197 patients:

- Row names are patient identifiers
- Column names are official HGNC gene symbols

Each element of the matrix is coded:

-2, homozygous deletions
-1, hemizygous deletions
0, neutral
1, gain
2, amplifications

Source

https://www.synapse.org/#!Synapse:syn300013
cna.logr

Description
A dataset containing part of the The Cancer Genome Atlas acute myeloid leukemia Affymetrix SNP6.0 array copy number alteration log2 ratios

Usage
cna.logr

Format
A matrix containing the copy number log2 ratios of 454 genes in 197 patients:
- Row names are patient identifiers
- Column names are official HGNC gene symbols

Source
https://www.synapse.org/#!Synapse:syn300013

ConvertXseqOutput
Convert xseq output to a data.frame

Description
Convert xseq output to a data.frame

Usage
ConvertXseqOutput(posterior)

Arguments
posterior The posterior probabilities of mutations and mutated genes, output from InferXseqPosterior

Value
A data.frame with sample, gene, probability of individual mutations and the probabilities of individual mutated genes
A mixture modelling approach to estimate whether a gene is expressed in a study given RNA-seq gene expression data

Usage

```r
EstimateExpression(expr, show.plot = FALSE, loglik = TRUE,
                    xlab = "Expression", ylab = "Density", ...)
```

Arguments

- `expr`: A matrix of RNA-seq gene expression values where each row corresponds to a patient and each column is a gene. Typically the expression of each gene is the log2 transformed RSEM value.
- `show.plot`: Logical, specifying whether to plot results
- `loglik`: Logical, whether plot the log-likelihoods
- `xlab`: `xlab` of the plot
- `ylab`: `ylab` of the plot
- `...`: Arguments for plotting

Value

A weight vector representing whether individual genes are expressed in the study

Examples

```r
data(expr)
weight = EstimateExpression(expr)
```

TCGA AML SNP6.0 gene expression data

Description

A dataset containing part of the The Cancer Genome Atlas acute myeloid leukemia RNA-seq gene expression data

Usage

`expr`
FilterNetwork

Format
A matrix containing the expression of 454 genes in 197 patients:
- Row names are patient identifiers
- Column names are official HGNC gene symbols

Source
https://www.synapse.org/

Description
Filter network

Usage
FilterNetwork(net, weight, min.weight = 0.8, min.conn.strength = 0.4,
min.num.conn = 5, max.num.conn = 50, remove.self.connection = TRUE,
double = FALSE)

Arguments
net
List, a gene interaction network
weight
The weights of genes, could from the function EstimateExpression
min.weight
Filter the connected genes with weights less than min.weight
min.conn.strength
The minimum gene connection strength
min.num.conn
The minimum number of connections required for a gene to be considered for
trans-analysis
max.num.conn
Only keep the top max.conn genes
remove.self.connection
Logical, whether removing self-connections or not
double
Logical, specifying whether debug information should be printed

Value
The filtered network

Examples
data(net)
net.filt = FilterNetwork(net)
GetExpressionDistribution

Get the conditional distributions for a set of genes

Description

Get the conditional distributions for a set of genes

Usage

GetExpressionDistribution(expr, mut = NULL, cna.call = NULL, gene = NULL,
type = "student", show.plot = FALSE)

Arguments

expr A matrix of gene expression values where each row corresponds to a patient and
each column is a gene.

mut A data.frame of mutations. The data.frame should have three columns of char-
acters: sample, hgnc_symbol, and variant_type. The variant_type column cat be either "HOMD", "HLAMP", "MISSENSE", "NONSENSE", "FRAMESHIFT", "INFRAME", "SPLICE", "NONSTOP", "STARTGAINED", "SYNONYMOUS", "OTHER", "FUSION", "COMPLEX".

cna.call A matrix containing the copy number calls, where each element is coded:
• -2, homozygous deletions
• -1, hemizygous deletions
• 0, neutral
• 1, gain
• 2, amplifications

gene A character vector of official HGNC gene names

type Character, either Gaussian ("gauss") or Student ("student")

show.plot Logical, specifying whether to plot the fitted model

Value

A list containing the fitted expression distributions
**ImputeKnn**

**ImputeKnn**

**Impute missing values (NAs) using K-nearest neighbour averaging**

**Description**

Impute missing values (NAs) using K-nearest neighbour averaging

**Usage**

ImputeKnn(X, ratio = 0.5, ...)

**Arguments**

- **X**
  A matrix of real values where each row corresponds to a patient and each column is a gene.
- **ratio**
  The rows (columns) with more than (default 50%) of missing values are removed
- **...**
  Arguments to be passed to impute.knn

**Value**

A matrix without NAs

**Examples**

data(expr)
expr.norm = ImputeKnn(expr)

**InferXseqPosterior**

**Learn xseq parameters given an initialized model**

**Description**

Learn xseq parameters given an initialized model

**Usage**

InferXseqPosterior(model, constraint, debug = FALSE)

**Arguments**

- **model**
  An xseq model.
- **constraint**
  A list of constraints on $\theta_{G|F}$.
- **debug**
  Logical, specifying whether debug information should be printed

**Value**

The posterior probabilities of latent variables in xseq
InitXseqModel

\textit{The datastructure to store the xseq models}

**Description**

The datastructure to store the xseq models

**Usage**

\begin{verbatim}
InitXseqModel(mut, expr, net, expr.dis, prior, cpd, gene, p.h, weight,
             cis = FALSE, debug = FALSE)
\end{verbatim}

**Arguments**

- **mut**: A data.frame of mutations. The data.frame should have three columns of characters: sample, hgnc_symbol, and variant_type. The variant_type column can be either "HOMD", "HLAMP", "MISSENSE", "NONSENSE", "FRAMESHIFT", "INFRAME", "SPLICE", "NONSTOP", "STARTGAINED", "SYNONYMOUS", "OTHER", "FUSION", "COMPLEX".

- **expr**: A matrix of gene expression values where each row corresponds to a patient and each column is a gene.

- **net**: A list of gene interaction networks.

- **expr.dis**: The fitted gene expression distributions, output from \texttt{get_expression_distribution}.

- **prior**: The prior for xseq, output from \texttt{set_xseq_prior}.

- **cpd**: A list of conditional probability tables for xseq, output from \texttt{set_xseq_prior}.

- **gene**: A character vector of gene names, default to all the genes with mutations.

- **p.h**: The down-regulation probability list of each gene connected to a mutated gene, typically from running \texttt{learn_xseq_parameter} on a discovery dataset.

- **weight**: The weight list of each gene connected to a mutated gene, typically from running \texttt{learn_xseq_parameter} on a discovery dataset.

- **cis**: Logical, cis or trans analysis.

- **debug**: Logical, whether to output debug information.

**Value**

A xseq model.
LearnXseqParameter

Learn xseq parameters given an initialized model

Description
Learn xseq parameters given an initialized model

Usage
LearnXseqParameter(model, constraint, iter.max = 20, threshold = 1e-05,
cis = FALSE, debug = FALSE, show.plot = TRUE)

Arguments

model | An xseq model
constraint | A list of constraints on $\theta_{G|F}$.
iter.max | Maximum number of iterations in learning xseq parameters
threshold | The threshold to stop learning parameters
cis | Logical, cis-analysis or trans-analysis
debug | Logical, specifying whether debug information should be printed
show.plot | Logical, specifying whether to plot the Log-Likelihoods

Value
A list including the learned xseq model

mut

TCGA AML somatic mutation data

Description
A dataset containing the The Cancer Genome Atlas acute myeloid leukemia somatic mutation data

Usage
mut

Format
A data frame with 2311 rows and 12 variables:

- sample. character, patient identifier
- hgnc_symbol. character, official HGNC gene symbols
- variant_type. character, mutation type, can be either "HOMD", "HLAMP", "MISSENSE", "NONSENSE", "FRAMESHIFT", "INFRAME", "SPLICE", "NONSTOP", "STARTGAINED", "SYNONYMOUS", "OTHER", "FUSION", "COMPLEX"
- ...

**Source**

https://www.synapse.org/#!Synapse:syn1729383

---

**net**

*A networks containing gene associations*

---

**Description**

A list of gene interactions

**Usage**

net

**Format**

A list with 16 elements:

- List names are official HGNC gene symbols
- Each element of the list is a vector, and the name of the vector is an official HGNC gene symbol. Each element of the vector is a real number between 0 and 1, representing the association strength between two genes. The names of the elements of the vector are also official HGNC gene symbols.

---

**NormExpr**

*Remove the cis-effects of copy number alterations on gene expression*

---

**Description**

Remove the cis-effects of copy number alterations on gene expression

**Usage**

NormExpr(expr, cna.logr, gene, type = "gp", debug = FALSE, show.plot = FALSE, show.norm = TRUE)

**Arguments**

- **expr**
  A matrix of gene expression values where each row corresponds to a patient and each column is a gene.
- **cna.logr**
  A matrix of copy number alterations log2 ratio where each row corresponds to a patient and each column is a gene.
- **gene**
  A character vector of gene HGNC symbols, default for all genes with both gene expression and copy number log2 ratio data.
PlotRegulationHeatmap

<table>
<thead>
<tr>
<th>type</th>
<th>A character, either Gaussian process regression (&quot;gp&quot;) or support vector machine regression (&quot;svm&quot;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>debug</td>
<td>Logical, specifying whether debug information should be printed</td>
</tr>
<tr>
<td>show.plot</td>
<td>Logical, specifying whether to plot the original expression and the normalized expression for a gene</td>
</tr>
<tr>
<td>show.norm</td>
<td>Logical, specifying whether to plot the expression of a gene after normalization, only used when show.plot = TRUE.</td>
</tr>
</tbody>
</table>

Value

The normalized expression matrix

Examples

```r
data(cna.logr, expr)
expr.norm = NormExpr(cna.logr, expr, gene="PTEN")
```

### PlotRegulationHeatmap

Heatmap showing the connected genes’ dysregulation probabilities

Description

Heatmap showing the connected genes’ dysregulation probabilities

Usage

```r
PlotRegulationHeatmap(gene, posterior, mut, subtype = NULL,
                        main = "in_Cancer", ...)
```

Arguments

- **gene**: A character vector of gene names
- **posterior**: The xseq posteriors, output of InferXseqPosterior or LearnXseqParameter
- **mut**: A data.frame of mutations. The data.frame should have at least three columns of characters: sample, hgnc_symbol, and variant_type. The variant_type column can be either "HOMD", "HLAMP", "MISSENSE", "NONSENSE", "FRAMESHIFT", "INFRAME", "SPLICE", "NONSTOP", "STARTGAINED", "SYNONYMOUS", "OTHER", "FUSION", "COMPLEX".
- **subtype**: A vector representing a character of each patient, e.g., subtype
- **main**: The heatmap title
- **...**: Other parameters passed to heatmap.2
QuantileNorm

Quantile normalize a matrix

**Description**
Quantile normalize a matrix

**Usage**
QuantileNorm(X)

**Arguments**

- **X**
  
  A matrix of real values where each row corresponds to a patient and each column is a gene

**Value**

The normalized matrix of X

**Examples**

data(expr)
expr.quantile = QuantileNorm(expr)

---

SetXseqPrior

Set model parameter priors

**Description**
Set model parameter priors

**Usage**
SetXseqPrior(mut, expr.dis, net, regulation.direction = TRUE, cis = TRUE, mut.type = "loss", ...)

**Arguments**

- **mut**
  
  A data.frame of mutations. The data.frame should have three columns of characters: sample, hgnc_symbol, and variant_type. The variant_type column can be either "HOMD", "HLAMP", "MISSENSE", "NONSENSE", "FRAMESHIFT", "INFRAME", "SPLICE", "NONSTOP", "STARTGAINED", "SYNONYMOUS", "OTHER", "FUSION", "COMPLEX".

- **expr.dis**
  
  A list, the outputs from calling GetExpressionDistribution

- **net**
  
  A list of gene interactions
SetXseqPrior

- **regulation.direction**: Logical, whether considering the directionality, i.e., up-regulation or down-regulation of genes, only used when cis=FALSE.
- **cis**: Logical, cis analysis or trans analysis
- **mut.type**: Character, only used when cis = TRUE, and can be either loss, gain or both
- **...**: Reserved for extension
Index

*Topic datasets
  cna.call, 2
  cna.logr, 3
  expr, 4
  mut, 9
  net, 10

cna.call, 2
cna.logr, 3
ConvertXseqOutput, 3

EstimateExpression, 4
expr, 4

FilterNetwork, 5

GetExpressionDistribution, 6

ImputeKnn, 7
InferXseqPosterior, 7
InitXseqModel, 8

LearnXseqParameter, 9
mut, 9

net, 10
NormExpr, 10

PlotRegulationHeatmap, 11

QuantileNorm, 12

SetXseqPrior, 12