**Package ‘DegNorm’**

April 1, 2024

<table>
<thead>
<tr>
<th>Type</th>
<th>Package</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td>DegNorm: degradation normalization for RNA-seq data</td>
</tr>
<tr>
<td><strong>Version</strong></td>
<td>1.12.0</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>2022-06-03</td>
</tr>
<tr>
<td><strong>Author</strong></td>
<td>Bin Xiong and Ji-Ping Wang</td>
</tr>
<tr>
<td><strong>Maintainer</strong></td>
<td>Ji-Ping Wang &lt;<a href="mailto:jzwang@northwestern.edu">jzwang@northwestern.edu</a>&gt;</td>
</tr>
<tr>
<td><strong>biocViews</strong></td>
<td>RNASeq, Normalization, GeneExpression, Alignment,Coverage, DifferentialExpression, BatchEffect,Software,Sequencing, ImmunoOncology, QualityControl, DataImport</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>This package performs degradation normalization in bulk RNA-seq data to improve differential expression analysis accuracy.</td>
</tr>
<tr>
<td><strong>License</strong></td>
<td>LGPL (&gt;= 3)</td>
</tr>
<tr>
<td><strong>Depends</strong></td>
<td>R (&gt;= 4.0.0), methods</td>
</tr>
<tr>
<td><strong>Imports</strong></td>
<td>Rcpp (&gt;= 1.0.2), GenomicFeatures, parallel, foreach, S4Vectors, doParallel, Rsamtools (&gt;= 1.31.2), GenomicAlignments, heatmaply, data.table, stats, ggplot2, GenomicRanges, IRanges, plyr, plotly, utils, viridis</td>
</tr>
<tr>
<td><strong>LinkingTo</strong></td>
<td>Rcpp, RcppArmadillo, S4Vectors, IRanges</td>
</tr>
<tr>
<td><strong>NeedsCompilation</strong></td>
<td>yes</td>
</tr>
<tr>
<td><strong>Suggests</strong></td>
<td>knitr, rmarkdown, formatR</td>
</tr>
<tr>
<td><strong>VignetteBuilder</strong></td>
<td>knitr</td>
</tr>
<tr>
<td><strong>BugReports</strong></td>
<td><a href="https://github.com/jipingw/DegNorm/issues">https://github.com/jipingw/DegNorm/issues</a></td>
</tr>
<tr>
<td><strong>git_url</strong></td>
<td><a href="https://git.bioconductor.org/packages/DegNorm">https://git.bioconductor.org/packages/DegNorm</a></td>
</tr>
<tr>
<td><strong>git_branch</strong></td>
<td>RELEASE_3_18</td>
</tr>
<tr>
<td><strong>git_last_commit</strong></td>
<td>405e5cb</td>
</tr>
<tr>
<td><strong>git_last_commit_date</strong></td>
<td>2023-10-24</td>
</tr>
<tr>
<td><strong>Repository</strong></td>
<td>Bioconductor 3.18</td>
</tr>
<tr>
<td><strong>Date/Publication</strong></td>
<td>2024-04-01</td>
</tr>
</tbody>
</table>
DegNorm: degradation normalization for RNA-seq data

Description

DegNorm is an R package for degradation normalization for bulk RNA-seq data. DegNorm, short for degradation normalization, is a bioinformatics pipeline designed to correct for bias due to the heterogeneous patterns of transcript degradation in RNA-seq data.

Details

DegNorm is a data-driven approach for RNA-Seq normalization resulting in the adjusted read count matrix. This adjustment applies to each gene within each sample, accounting for sample- and gene-specific degradation bias while simultaneously controlling for the sequencing depth. The algorithm at the center of DegNorm is the rank-one over-approximation of a gene’s coverage score matrix, which is comprised of the different samples’ coverage score curves along the transcript for each gene. For each gene, DegNorm estimates (1) an envelope function representing the ideal shape of the gene’s coverage curve when no degradation is present, and (2) scale factors for each sample (for said gene) that indicates the relative abundance of the gene within the sample.

functions: read_coverage_batch, degnorm, plot_coverage, plot_heatmap, plot_corr, plot_boxplot

Author(s)

Bin Xiong, Ji-Ping Wang

Maintainer: Ji-Ping Wang <jzwang@northwestern.edu>

References

DegNorm reference:

Xiong, B., Yang, Y., Fineis, F. Wang, J.-P., DegNorm: normalization of generalized transcript degradation improves accuracy in RNA-seq analysis, Genome Biology, 2019, 20:75
**coverage_res_chr21**  

Example **CoverageClass data**

---

**Description**

Example of **CoverageClass** data from **DegNorm** package. It is the output from **read_coverage_batch** function for human chromosome 21.

**Usage**

data(coverage_res_chr21)

**Format**

A **coverageClass** list of the following:

- **coverage**: a list of converage matrices for all genes within each sample
- **counts**: a data.frame of read counts for all genes within each sample.

**Examples**

data(coverage_res_chr21)  
summary_CoverageClass(coverage_res_chr21)

---

**degnorm**  

**Main function to perform degradation normalization.**

---

**Description**

degnorm calculates the degradation index score for each gene within each sample and return the degradation-normalized read counts.

**Usage**

degnorm(read_coverage, counts, iteration, loop, down_sampling=1, grid_size=10, cores=1)

**Arguments**

- **read_coverage**: a list of converage matrices, one per gene
- **counts**: dataframe of read counts, each row for one gene, and column for sample. The order and number of genes must match the order in read_coverage matrices.
- **iteration**: iteration number for degnorm algorithm. 5 is sufficient.
- **loop**: iteration number inside of nonnegative matrix factorization-over approximation. Default is 100.
DegNorm provides three functions for visualization gene-/sample-wise degradation.

### DegNorm-plot-functions

**Description**

DegNorm provides three functions for visualization gene-/sample-wise degradation.

**Down-sampling**

1 for yes (default) and 0 for no. If yes, average coverage score is calculated on a grid of size specified by grid_size argument. The new coverage matrix formed by the grid average score will be used for baseline selection. This increases the efficiency of algorithm while maintaining comparable accuracy.

**Grid size**

default size is 10 bp.

**Cores**

number of cores. Default number if 1. Users should input the maximum possible number of cores for efficiency.

**Value**

degnorm outputs a list of following objects:

- **counts**
  - a data.frame of read counts for each gene within each sample.

- **counts_normed**
  - a data.frame of degradation-normalized read counts for each gene within each sample.

- **DI**
  - a matrix of degradation index scores for each gene within each sample.

- **K**
  - normalizing scale factor for each gene within each sample after accounting for degradation normalization.

- **convergence**
  - convergence tag: 0 = degnorm was not done on this gene because smaller counts or too short length. 1 = degnorm was done with baseline selection. 2 = degnorm done without baseline selection because gene length (after filtering out low count regions)<200 bp. 3= baseline was found, but DI score is too large. 4 = baseline selection didn't converge.

- **envelop**
  - list of the envelop curves for all genes.

**Examples**

```r
## coverage_res_chr21 is a CoverageClass object from DegNorm Package.
data(coverage_res_chr21)
res_DegNorm = degnorm(read_coverage = coverage_res_chr21[[1]],
  counts = coverage_res_chr21[[2]],
  iteration = 2,
  down_sampling = 1,
  grid_size=10,
  loop = 20,
  cores=2)
```

```r
DegNorm-plot-functions

Degradation index (DI) score plot functions
```

```r
DegNorm-plot-functions

Degradation index (DI) score plot functions
```

```r
DegNorm-plot-functions

Degradation index (DI) score plot functions
```
**plot_coverage**

**Usage**

```r
plot_corr(DI)
plot_heatmap(DI)
plot_boxplot(DI)
```

**Arguments**

- **DI**: a matrix or data.frame of degradation index (DI) scores with each row corresponding to one gene and each column for a sample.

**Details**

`plot_corr` plots the correlation matrix of DI scores between samples. `plot_heatmap` plots the heatmap of DI scores. Left is plotted in descending order of average DI scores of genes where each row corresponds to one gene. In the right plot, DI scores were sorted within each sample and plotted in descending order. `plot_boxplot` plots the boxplot of DI scores by samples.

**Value**

These functions return a boxplot of DI scores by sample, a heatmap of DIS scores of all genes in all samples and a correlation plot of DI scores between samples respectively.

**Examples**

```r
## res_DegNorm_chr21 is degnorm output stored in sysdata.Rda
data(res_DegNorm_chr21)
plot_boxplot(res_DegNorm_chr21$DI)
plot_heatmap(res_DegNorm_chr21$DI)
plot_corr(res_DegNorm_chr21$DI)
```

---

**plot_coverage**  
*Coverage plot functions for DegNorm*

**Description**

`plot_coverage` plots the before- and after-degradation coverage curves

**Usage**

```r
plot_coverage(gene_name, coverage_output, degnorm_output, group=NULL, samples=NULL)
```

**Arguments**

- **gene_name**: the name of the gene whose coverage coverage to be plotted.
- **coverage_output**: CoverageClass object, the output from function `coverage_cal_batch`.
- **degnorm_output**: DegNormClass object, the output from function `DegNorm`.
read_coverage

小组 a vector of integers or character strings indicating the biological conditions of
    the samples. Coverage curves will be plotted in the same color for the same
group. Default is NULL. By default all curves will plotted in different colors.

samples a string vector for the subset of samples to be plotted. NULL means all samples
to be plotted. The length of samples must be of the same length of group if both
    specified.

Details
    plot_coverage outputs the coverage curves before- and after-degradation normalization.

Value
    The coverage curve before and after degradation normalization.

Examples
    ## gene named "SOD1", plot coverage curves
    data(coverage_res_chr21)
    data(res_DepDegNorm_chr21)
    plot_coverage(gene_name="SOD1", coverage_output=coverage_res_chr21,
        degnorm_output=res_DepDegNorm_chr21, group=c(0,1,1))

read_coverageFunction to calculate read coverage score for one bam file

Description
    This function judges whether bam file is single-end and paired-end, and generate bam file index
    if needed. It calls function paired_end_cov_by_ch or single_end_by_ch. It takes multiple-core
structure for parallel computing for efficiency.

Usage
    read_coverage(bam_file, all_genes, cores)

Arguments
    bam_file The name of the bam file.
    all_genes An GRangesList object. It’s the parsed genes annotation file from GTF file.
    cores number of cores to use.

Details
    This function judges whether bam file is single-end and paired-end, and generate bam file index if
    needed. It takes multiple-core structure for parallel computing for efficiency.
This function returns a coverageClass object. It contains a list of: (1) a list of coverage score for each gene in RLE format and (2) a dataframe for read counts.

See Also

`read_coverage_batch`
Examples

```r
## read bam file and gtf file from the package
bam_file_list <- list.files(path=system.file("extdata",package="DegNorm"),
  pattern=".bam$",full.names=TRUE)
gtf_file <- list.files(path=system.file("extdata",package="DegNorm"),
  pattern=".gtf$",full.names=TRUE)

# run read_coverage_batch to calculate read coverage curves and read counts
coverage_res=read_coverage_batch(bam_file_list, gtf_file, cores=2)
```

---

### Description

Example of DegNormClass data from DegNorm package. It is the output from degnorm function for human chromosome 21.

### Usage

```r
data("res_DegNorm_chr21")
```

### Format

A DegNormClass list of the following items:

- `counts` a data.frame of read counts for each gene within each sample.
- `counts_normed` a data.frame of degradation-normalized read counts for each gene within each sample.
- `DI` a matrix of degradation index scores for each gene within each sample.
- `K` normalizing scale factor for each gene within each sample after accounting for degradation normalization.
- `convergence` convergence tag; 0 = degnorm was not done on this gene because smaller counts or too short length. 1 = degnorm was done with baseline selection. 2 = degnorm done without baseline selection because gene length (after filtering out low count regions)<200 bp. 3 = baseline was found, but DI score is too large. 4 = baseline selection didn’t converge.
- `envelop` a list of the envelop curves for all genes.

### Examples

```r
data(res_DegNorm_chr21)
summary_DegNormClass(res_DegNorm_chr21)
```
### summary_CoverageClass

**Summary method for CoverageClass.**

**Description**

It prints a summary of the data objects contained in the list from `read_coverage_batch`.

**Usage**

```r
summary_CoverageClass(object)
```

**Arguments**

- `object` CoverageClass from `coderead_coverage_batch`.

**Value**

On-screen plot of summary of CoverageClass object.

**Examples**

```r
## Summary of coverage_cal_batch output (CoverageClass)
data(coverage_res_chr21)
summary_CoverageClass(coverage_res_chr21)
```

### summary_DegNormClass

**Summary method for DegNormClass.**

**Description**

It prints a summary of the data objects contained in the list from `degnorm` function.

**Usage**

```r
summary_DegNormClass(object)
```

**Arguments**

- `object` DegNormClass from `degnorm` function.

**Value**

On-screen summary of DegNormClass object.

**Examples**

```r
## Summary of degnorm output (DegNormClass)
data(res_DegNorm_chr21)
summary_DegNormClass(res_DegNorm_chr21)
```
Index

* RNA-seq, degradation, normalization
  DegNorm-package, 2
* datasets
  coverage_res_chr21, 3
  res_DegNorm_chr21, 8
* internal
  read_coverage, 6

coverage_res_chr21, 3
degnorm, 3
DegNorm-package, 2
DegNorm-plot-functions, 4

plot_boxplot (DegNorm-plot-functions), 4
plot_corr (DegNorm-plot-functions), 4
plot_coverage, 5
plot_heatmap (DegNorm-plot-functions), 4

read_coverage, 6, 7
read_coverage_batch, 7, 7
res_DegNorm_chr21, 8

summary_CoverageClass, 9
summary_DegNormClass, 9