Package 'awst'

January 31, 2025

Title Asymmetric Within-Sample Transformation

Version 1.15.0

Description We propose an Asymmetric Within-Sample Transformation (AWST) to regularize RNA-seq read counts and reduce the effect of noise on the classification of samples. AWST comprises two main steps: standardization and smoothing. These steps transform gene expression data to reduce the noise of the lowly expressed features, which suffer from background effects and low signal-to-noise ratio, and the influence of the highly expressed features, which may be the result of amplification bias and other experimental artifacts.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.1.1

URL https://github.com/drisso/awst

BugReports https://github.com/drisso/awst/issues

Imports stats, methods, SummarizedExperiment

Suggests airway, ggplot2, testthat, EDASeq, knitr, BiocStyle, RefManageR, sessioninfo, rmarkdown

biocViews Normalization, GeneExpression, RNASeq, Software, Transcriptomics, Sequencing, SingleCell

VignetteBuilder knitr

git_url https://git.bioconductor.org/packages/awst

git_branch devel

git_last_commit d18c8ed

git_last_commit_date 2024-10-29

Repository Bioconductor 3.21

Date/Publication 2025-01-31

Author Davide Risso [aut, cre, cph] (ORCID: <https://orcid.org/0000-0001-8508-5012>), Stefano Pagnotta [aut, cph] (ORCID: <https://orcid.org/0000-0002-8298-9777>)

Maintainer Davide Risso <risso.davide@gmail.com>

Contents

`awst` *Asymmetric Within-Sample Transformation*

Description

This function implements the asymmetric within-sample transformation described in Risso and Pagnotta (2019). The function includes two steps: a standardization step and a asymmetric winsorization step. See details.

Usage

```
## S4 method for signature 'matrix'
awst(x, poscount = FALSE, full_quantile = FALSE, sigma0 = 0.075, lambda = 13)

## S4 method for signature 'SummarizedExperiment'
awst(
  x,
  poscount = FALSE,
  full_quantile = FALSE,
  sigma0 = 0.075,
  lambda = 13,
  expr_values = "counts",
  name = "awst"
)
```

Arguments

`x`	a matrix of (possibly normalized) RNA-seq read counts or a 'SummarizedExperiment'.
`poscount`	a logical value indicating whether positive counts only should be used for the standardization step.
`full_quantile`	a logical value indicating whether the data have been normalized with the full-quantile normalization. In this case, computations can be sped up.
`sigma0`	a multiplicative constant to be applied to the smoothing function.
`lambda`	a parameter that controls the growth rate of the smoothing function.
`expr_values`	integer scalar or string indicating the assay that contains the matrix to use as input.

name — string specifying the name of the assay to be used to store the results of the transformation.

Details

The standardization step is based on a log-normal distribution of the high-intensity genes. Optionally, only positive counts can be used in this step (this option is especially useful for single-cell data). The winsorization step is controlled by two parameters, sigma0 and lambda, which control the growth rate of the winsorization function.

Value

if ‘x‘ is a matrix, it returns a matrix of transformed values, with genes in rows and samples in column. If ‘x‘ is a ‘SummarizedExperiment‘, it returns a ‘SummarizedExperiment‘ with the transformed value in the ‘name‘ slot.

Methods (by class)

- `matrix`: the input is a matrix of (possibly normalized) counts
- `SummarizedExperiment`: the input is a SummarizedExperiment with (possibly normalized) counts in one of its assays.

References

Risso and Pagnotta (2019). Within-sample standardization and asymmetric winsorization lead to accurate classification of RNA-seq expression profiles. Manuscript in preparation.

Examples

```
x <- matrix(data = rpois(100, lambda=5), ncol=10, nrow=10)
awst(x)
```

---

gene_filter *Gene filtering based on heterogeneity*

Description

This function filters out genes that show a low heterogeneity, as measured by Shannon's entropy.

Usage

```
## S4 method for signature 'matrix'
gene_filter(
  x,
  from = min(x, na.rm = TRUE),
  to = max(x, na.rm = TRUE),
  nBins = 20,
```

```
  heterogeneity_threshold = 0.1
)

## S4 method for signature 'SummarizedExperiment'
gene_filter(
  x,
  from = min(assay(x, awst_values), na.rm = TRUE),
  to = max(assay(x, awst_values), na.rm = TRUE),
  nBins = 20,
  heterogeneity_threshold = 0.1,
  awst_values = "awst"
)
```

Arguments

x	a matrix of transformed gene expression counts (typically the results of awst).
from	the minimum value from which to start binning data.
to	the maximum value for the binning of the data.
nBins	the number of bins.
heterogeneity_threshold	the trheshold used for the filtering.
awst_values	integer scalar or string indicating the assay that contains the awst-transformed values to use as input.

Details

Shannon's entropy is computed on the categorized data after AWST transformation. Those genes that show a lower entropy than the predefined threshold are deemed to carry too low information to be useful for the classification of the samples, and are hence removed.

Value

if 'x' is a matrix, it returns a filtered matrix. If 'x' is a 'SummarizedExperiment', it returns a filtered 'SummarizedExperiment'

Methods (by class)

- matrix: the input is a matrix of awst-transformed values.
- SummarizedExperiment: the input is a SummarizedExperiment with awst-transformed values in one of its assays.

References

Risso and Pagnotta (2019). Within-sample standardization and asymmetric winsorization lead to accurate classification of RNA-seq expression profiles. Manuscript in preparation.

Examples

```
set.seed(222)
x <- matrix(rpois(75, lambda=5), ncol=5, nrow=15)
a <- awst(x)
gene_filter(a)
```

Index