Package 'MoleculeExperiment'

June 1, 2023

Title Prioritising a molecule-level storage of Spatial Transcriptomics Data

Version 1.0.1

Description MoleculeExperiment contains functions to create and work with objects from the new MoleculeExperiment class. We introduce this class for analysing molecule-based spatial transcriptomics data (e.g., Xenium by 10X, Cosmx SMI by Nanostring, and Merscope by Vizgen). This allows researchers to analyse spatial transcriptomics data at the molecule level, and to have standardised data formats accross vendors.

License MIT + file LICENSE

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.3

URL https://github.com/SydneyBioX/MoleculeExperiment

BugReports https://github.com/SydneyBioX/MoleculeExperiment/issues

Imports SpatialExperiment, Matrix, purrr, sp, data.table, dplyr (>= 1.1.1), magrittr, rjson, utils, methods, terra, ggplot2, rlang

Suggests knitr, BiocStyle, testthat (>= 3.0.0)

VignetteBuilder knitr

biocViews DataImport, DataRepresentation, Infrastructure, Software, Spatial, Transcriptomics

Config/testthat/edition 3

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

git_branch RELEASE_3_17

git_last_commit b21c085

git_last_commit_date 2023-05-08

Date/Publication 2023-06-01

Author Bárbara Zita Peters Couto [aut],
Nicholas Robertson [aut],
Ellis Patrick [aut],
Shila Ghazanfar [aut, cre]

Maintainer Shila Ghazanfar <shazanfar@gmail.com>

R topics documented:

`accessors` *Accessor functions to work with MoleculeExperiment objects*

Description

Accessor functions to work with MoleculeExperiment objects

Usage

```
## S4 method for signature 'MoleculeExperiment'
molecules(object, assayName = "detected", flatten = FALSE)

## S4 method for signature 'MoleculeExperiment'
boundaries(object, assayName = NULL, flatten = FALSE)

## S4 method for signature 'MoleculeExperiment'
features(object, assayName = "detected")

## S4 method for signature 'MoleculeExperiment'
segmentIDs(object, assayName = NULL)

## S4 replacement method for signature 'MoleculeExperiment'
molecules(object, assayName = NULL) <- value

## S4 replacement method for signature 'MoleculeExperiment'
boundaries(object, assayName = NULL) <- value
```

Arguments

object	The MoleculeExperiment to access.
assayName	Character string specifying the name of the assay from which to retrieve or set information in the slot of interest.
flatten	Logical value specifying whether to flatten the ME list into a data.frame or not. Defaults to FALSE.
value	New value to be added to the slot and assay of interest.

Value

A MoleculeExperiment object slot.

getters

Accessor functions to get data from the MoleculeExperiment object. These include:

- molecules() to retrieve information from the molecules slot.
- boundaries() to retrieve information from the boundaries slot.
- features() to retrieve feature names from the molecules slot.
- segmentIDs() to retrieve segment ids from the boundaries slot.

setters

The molecules<- setter accesses the molecules slot, whereas the boundaries slot can be accessed with boundaries<-.

Examples

```
# get example data
repoDir <- system.file("extdata", package = "MoleculeExperiment")
repoDir <- paste0(repoDir, "/xenium_V1_FF_Mouse_Brain")
me <- readXenium(repoDir,
                 keepCols = "essential",
                 addBoundaries = "cell")

# get insight into molecules slot
showMolecules(me)

# for developers, use molecules() getter
# expect a large output from call below
# molecules(me)
# alternatively, return rectangular data structure with flatten = TRUE
molecules(me, assayName = "detected", flatten = TRUE)

# get insight into boundaries slot
showBoundaries(me)

# for developers, use boundaries() getter
# expect a large output from call below
```

```
# boundaries(me, assayName = "cell")
# alternatively, return rectangular data structure with flatten = TRUE
boundaries(me, assayName = "cell", flatten = TRUE)

# features() getter
features(me)

# segmentIDs() getter
segmentIDs(me, "cell")

# setter example
# read in and standardise nucleus boundaries too
nucleiMEList <- readBoundaries(dataDir = repoDir,
                               pattern = "nucleus_boundaries.csv",
                               segmentIDCol = "cell_id",
                               xCol = "vertex_x",
                               yCol = "vertex_y",
                               keepCols = "essential",
                               boundariesAssay = "nucleus",
                               scaleFactorVector = 1)

# use `boundaries<-` setter to add nucleus boundaries to the boundaries slot
boundaries(me, "nucleus") <- nucleiMEList
me
```

`countMolecules` *Count molecules per region of interest (e.g., cell)*

Description

This function takes the information from the molecules and boundaries slot, and counts the molecules per region of interest. Its input is a MoleculeExperiment object, and its output a SpatialExperiment object. That way, if one is interested in doing downstream analyses at the cell level, one can do so.

Usage

```
countMolecules(
  object,
  moleculesAssay = "detected",
  segmentationInfo = "boundaries",
  boundariesAssay = "cell",
  matrixOnly = FALSE
)
```

Arguments

`object`	MoleculeExperiment object containing both the transcript data as well as the boundaries data. I.e., the "molecules" and "boundaries" slots need to be filled. See MoleculeExperiment() for more information.

`moleculesAssay` Character string naming the list of the molecules slot from which transcript information should be retrieved from. The default is the detected transcript data that is read in when creating a MoleculeExperiment object. It is possible to change it to another mode, e.g., "high_threshold" will access the transcript information that has been stored in the "high_threshold" element of the list in the molecules slot.

`segmentationInfo` Character string specifying the type of segmentation information available. Can be either "boundaries" or "masks". Currently, only the "boundaries" information is supported.

`boundariesAssay` Character string naming the list of the boundaries slot from which boundary information should be retrieved from. For example, for counting transcripts per cell, the list containing the cell boundaries (e.g., "cell") should be selected.

`matrixOnly` Logical value indicating whether to return a matrix of the counted molecules per segment (e.g., cell). Is FALSE by default, i.e., the default output is a SpatialExperiment object.

Value

A SpatialExperiment object derived from a MoleculeExperiment object. Alternatively, a matrix with the counted molecules per segment.

Examples

```
repoDir <- system.file("extdata", package = "MoleculeExperiment")
repoDir <- paste0(repoDir, "/xenium_V1_FF_Mouse_Brain")
me <- readXenium(repoDir,
                 keepCols = "essential")

spe <- countMolecules(me)
spe
```

`dataframeToMEList` *Convert a transcripts or boundaries file to the ME list format*

Description

The goal of this function is to standardise transcripts and boundaries files for input to a MoleculeExperiment object.

Usage

```
dataframeToMEList(
  df,
  dfType = NULL,
  assayName = NULL,
```

```
  sampleCol = "sample_id",
  factorCol = NULL,
  xCol = "x_location",
  yCol = "y_location",
  keepCols = "essential",
  scaleFactor = 1
)
```

Arguments

df	A data.frame containing the transcripts information or the boundaries information. NOTE: this dataframe should, at a minimum, have the following 4 columns: sample_id, factorCol (e.g., feature_name in transcripts, or cell_id in boundaries), x_location and y_location.
dfType	Character string specifying contents of the dataframe. Can be either "transcripts" or "boundaries".
assayName	Character string specifying the name with which to identify the information later on in an ME object.
sampleCol	Character string specifying the name of the column with the sample id.
factorCol	Character string specifying the name of the column with the factors with which to group the data in the lists. When working with molecules, this column would be e.g., "feature_name" in xenium. When working with boundaries, this column would be e.g., "cell_id" in xenium.
xCol	Character string specifying the name of the column with global x coordinates.
yCol	Character string specifying the name of the column with global y coordinates.
keepCols	Character string which can be either "essential" or "all". If "essential", the function will only work with the x and y location information.
scaleFactor	Integer specifying the scale factor by which to change the scale of the x and y locations (e.g., to change from pixel to micron). The default value is 1.

Value

A list with the format required to input it into slots of a MoleculeExperiment object.

Examples

```
moleculesDf <- data.frame(
    sample_id = rep(c("sample1", "sample2"), times = c(30, 20)),
    features = rep(c("gene1", "gene2"), times = c(20, 30)),
    x_coords = runif(50),
    y_coords = runif(50)
)

moleculesMEList <- dataframeToMEList(moleculesDf,
                                     dfType = "transcripts",
                                     assayName = "detected",
                                     sampleCol = "sample_id",
                                     factorCol = "features",
```

```
                                    xCol = "x_coords",
                                    yCol = "y_coords")

moleculesMEList
```

`MoleculeExperiment-class` *MoleculeExperiment class: An S4 class container to store imaging-based spatial transcriptomics data.*

Description

This class enables the analysis of imaging-based ST data at the molecule level, and standardises data across vendors. The aim of this class is to facilitate ST data integration and comparison and, importantly, facilitate common analytical and visualisation workflows.

Usage

```
MoleculeExperiment(molecules, boundaries = NULL)
```

Arguments

`molecules`	Detected transcripts information in a standardised ME list format, as is generated by dataframeToMEList() and readMolecules() functions.
`boundaries`	Slot with boundary information in a standardised ME list format, as is generated by dataframeToMEList() and readBoundaries() functions.

Value

A MoleculeExperiment object

Slots

`molecules` Slot containing information about the detected transcripts. This slot is designed as a list of lists, where each sample contains a list of tibbles with information for each gene. The basic information required for this slot are the gene names of the transcripts, as well as their x and y locations.

`boundaries` Slot containing the boundaries defining each segmented cell. The slot is designed as a list of lists, where each sample contains a list of tibbles for each cell, consisting of the x and y coordinates of the polygon vertices defining the cell boundary.

Examples

```
# creating a simple ME object from toy data
moleculesDf <- data.frame(
    sample_id = rep(c("sample1", "sample2"), times = c(30, 20)),
    features = rep(c("gene1", "gene2"), times = c(20, 30)),
    x_coords = runif(50),
    y_coords = runif(50)
)
boundariesDf <- data.frame(
    sample_id = rep(c("sample1", "sample2"), times = c(16, 6)),
    cell_id = rep(c("cell1", "cell2", "cell3", "cell4",
                    "cell1", "cell2"),
                  times = c(4, 4, 4, 4, 3, 3)),
    vertex_x = rnorm(22),
    vertex_y = rnorm(22)
)
moleculesMEList <- dataframeToMEList(moleculesDf,
                                     dfType = "transcripts",
                                     assayName = "detected",
                                     sampleCol = "sample_id",
                                     factorCol = "features",
                                     xCol = "x_coords",
                                     yCol = "y_coords")

boundariesMEList <- dataframeToMEList(boundariesDf,
                                      dfType = "boundaries",
                                      assayName = "cell",
                                      sampleCol = "sample_id",
                                      factorCol = "cell_id",
                                      xCol = "vertex_x",
                                      yCol = "vertex_y")

toyME <- MoleculeExperiment(molecules = moleculesMEList,
                            boundaries = boundariesMEList)

toyME
```

---

plotting-functions *Plotting functions for SpatialUtils*

Description

A set of ggplot functions to build customized plots for imaging based spatial transcriptomics data.

Usage

```
ggplot_me()
```

```
geom_point_me(me, assayName = "detected", byColour = NULL, ...)

geom_polygon_me(me, assayName = "cell", byFill = NULL, ...)
```

Arguments

me	MoleculeExperiment object.
assayName	Character string specifying name of assay from which to get data.
byColour	Character string specifying the column name to colour by.
byFill	Character string specifying the column name to fill by.

Value

A plot with transcripts and/or segmentation information for imaging based spatial transcriptomics data.

Examples

```
repoDir <- system.file("extdata", package = "MoleculeExperiment")
repoDir <- paste0(repoDir, "/xenium_V1_FF_Mouse_Brain")
me <- readXenium(repoDir,
                 keepCols = "essential",
                 addBoundaries = c("cell", "nucleus"))

g = ggplot_me() +
        geom_polygon_me(me, byFill = "segment_id", colour = "black") +
        geom_point_me(me, byColour = "feature_name", size = 0.1) +
        geom_polygon_me(me, assayName = "nucleus", fill = NA, colour = "red")
g
```

---

readBoundaries *Read in csv boundary information and convert to ME list format.*

Description

This function reads in csv boundary files and converts them to the ME list format, so that they can be added to an ME object later on. To account for different coordinate scales possible being used by the boundary versus transcript information, this function scales the coordinate values of the boundaries to match the unit of the detected transcript locations. The various arguments offer flexibility to standardise data from different molecule-based ST technologies into the ME list format.

Usage

```
readBoundaries(
  dataDir,
  pattern = NULL,
  segmentIDCol = NULL,
  xCol = NULL,
```

```
  yCol = NULL,
  keepCols = "essential",
  boundariesAssay = NULL,
  scaleFactorVector = 1
)
```

Arguments

dataDir	Path of the directory containing the boundary csv files.
pattern	Character string specifying the unique pattern with which to identify the files of interest in the directory. This is useful to work with multiple samples. Defaults to NULL.
segmentIDCol	Character string specifying the name of the column containing the segment IDs. Defaults to NULL.
xCol	Character string specifying the name of the column containing the x coordinates of the vertices defining the boundaries. Defaults to NULL.
yCol	Character string specifying the name of the column containing the y coordinates of the vertices defining the boundaries. Defaults to NULL.
keepCols	Character string specifying which columns to keep. Defaults to "essential". The other option is to select "all", or custom columns by specifying their names in a vector.
boundariesAssay	Character string specifying the name with which to identify the boundary data in the ME object later on. Defaults to NULL.
scaleFactorVector	Vector containing the scale factor/s with which to change the coordinate data from pixel to micron. It can be either a single integer, or multiple scale factors for the different samples. The default value is 1.

Value

An ME list containing the boundary information. This can be used as input to the boundaries slot of an ME object.

Examples

```
repoDir <- system.file("extdata", package = "MoleculeExperiment")
repoDir <- paste0(repoDir, "/xenium_V1_FF_Mouse_Brain")
nucleiMEList <- readBoundaries(dataDir = repoDir,
                               pattern = "nucleus_boundaries.csv",
                               segmentIDCol = "cell_id",
                               xCol = "vertex_x",
                               yCol = "vertex_y",
                               keepCols = "essential",
                               boundariesAssay = "nucleus",
                               scaleFactorVector = 1)
nucleiMEList
```

readCosmx *Read in Cosmx data (Nanostring) as an ME object.*

Description

This function is a wrapper around the readMolecules function. Note that it can currently only create a simple ME object with the molecules slot filled. Boundary information is not handled yet.

Usage

```
readCosmx(dataDir, keepCols = "essential")
```

Arguments

dataDir	Character string specifying the directory with the Cosmx output files.
keepCols	Character string specifying which columns to keep. Defaults to "essential". The other option is to select "all", or custom columns by specifying their names in a vector.

Value

A MoleculeExperiment object

Examples

```
repoDir <- system.file("extdata", package = "MoleculeExperiment")
repoDir <- paste0(repoDir, "/nanostring_Lung9_Rep1")
meCosmx <- readCosmx(repoDir,
                     keepCols = "essential")
meCosmx
```

readMerscope *Read in Merscope data to an ME object*

Description

Reads in Merscope data (Vizgen) from a directory, and standardises it into a MoleculeExperiment object. It is essentially a wrapper around the function readMolecules(). Note that this function can currently only create a simple ME object with the molecules slot filled. Boundary information cannot be handled yet.

Usage

```
readMerscope(dataDir, keepCols = "essential")
```

Arguments

dataDir	Character string specifying the directory with the Cosmx output files.
keepCols	Vector of characters specifying the columns of interest from the transcripts file. "essential" selects columns with gene names, x and y locations. "all" will select all columns. Alternatively, specific colums of interest can be selected by specifying them as characters in a vector. Note that this personalised vector needs to contain the essential columns.

Value

A MoleculeExperiment object

Examples

```
repoDir <- system.file("extdata", package = "MoleculeExperiment")
repoDir <- paste0(repoDir, "/vizgen_HumanOvarianCancerPatient2Slice2")
meMerscope <- readMerscope(repoDir,
                           keepCols = "essential")
meMerscope
```

---

readMolecules *Read in detected transcripts file/s into a MoleculeExperiment object*

Description

A function to standardise transcripts.csv files across different molecule- based ST technologies, and store them into an ME object. It is technology agnostic, so it is accompanied with wrappers for the specific technologies (e.g., see readXenium).

Usage

```
readMolecules(
  dataDir,
  pattern = NULL,
  featureCol = NULL,
  xCol = NULL,
  yCol = NULL,
  keepCols = "essential",
  moleculesAssay = NULL,
  scaleFactorVector = 1
)
```

Arguments

dataDir	Character string specifying the directory with the file/s containing detected transcripts for different runs/samples.
pattern	Character string specifying the pattern with which to find the transcripts files. For example, in Xenium data, the pattern would be "transcripts.csv". In contrast, in Cosmx data, the pattern would be "tx_file".
featureCol	Character string specifying the name of the column with feature names. For example, "feature_name" in xenium transcripts.csv files.
xCol	Character string specifying the name of the column with the x locations of the transcripts.
yCol	Character string specifying the name of the column with the y locations of the transcripts.
keepCols	Vector of characters specifying the columns of interest from the transcripts file. "essential" selects columns with gene names, x and y locations. "all" will select all columns. Alternatively, specific colums of interest can be selected by specifying them as characters in a vector. Note that this personalised vector needs to contain the essential columns.
moleculesAssay	Character string specifying the name of the list in which the transcript information is going to be stored in the molecules slot. The default name is "detected", as we envision that a MoleculeExperiment will usually be created with raw detected transcript information.
scaleFactorVector	Vector containing the scale factor/s with which to change the coordinate data from pixel to micron. It can be either a single integer, or multiple scale factors for the different samples. The default value is 1.

Value

A simple MoleculeExperiment object with a filled molecules slot.

Examples

```
repoDir <- system.file("extdata", package = "MoleculeExperiment")
repoDir <- paste0(repoDir, "/xenium_V1_FF_Mouse_Brain")
simple_me <- readMolecules(repoDir,
                           pattern = "transcripts.csv",
                           featureCol = "feature_name",
                           xCol = "x_location",
                           yCol = "y_location",
                           keepCols = "essential")
simple_me
```

`readXenium` *Read in Xenium data into a MoleculeExperiment object*

Description

Function to read in, and standardise, Xenium output data into an ME object. Detected transcripts files are required. Additionally, it is also possible to read in boundary files ("cell", "nuclei", or both). This function is a wrapper around readMolecules and readBoundaries functions.

Usage

```
readXenium(dataDir, keepCols = "essential", addBoundaries = "cell")
```

Arguments

`dataDir`	Character string specifying the directory with the xenium output files.
`keepCols`	Vector of characters specifying the columns of interest from the transcripts file and boundaries file. Can be "all" or "essential". "essential" selects columns with gene names, x and y locations in the transcripts file; "essential" selects columns with cell ids, and x and y locations for the vertices defining the boundaries in the boundaries file.
`addBoundaries`	Vector with which to specify the names of the boundary assays to be added to the me object. Can be "cell", "nucleus", both, or NULL. The latter will lead to the creation of a simple ME object with just the molecules slot filled.

Value

A MoleculeExperiment object containing xenium data.

Examples

```
repoDir <- system.file("extdata", package = "MoleculeExperiment")
repoDir <- paste0(repoDir, "/xenium_V1_FF_Mouse_Brain")

me <- readXenium(repoDir,
                 keepCols = "essential")
me
```

summarization *Summarization methods to get insights into a MoleculeExperiment object*

Description

The following methods are useful to get quick view of the contents in a MoleculeExperiment object. For example, showMolecules and showBoundaries summarise the large nested ME list of lists in the molecules and boundaries slots. nFeatures and nTranscripts get the numbers of features or transcripts, respectively. They can do so across all samples, or per sample.

Usage

```
## S4 method for signature 'MoleculeExperiment'
showMolecules(object)

## S4 method for signature 'MoleculeExperiment'
showBoundaries(object)

## S4 method for signature 'MoleculeExperiment'
nFeatures(object, assayName = "detected", perSample = FALSE)

## S4 method for signature 'MoleculeExperiment'
nTranscripts(object, assayName = "detected", perSample = FALSE)
```

Arguments

object	Name of MoleculeExperiment object of interest.
assayName	Character string specifying the name of the assay from which to view a summary of the contents.
perSample	Logical value specifying whether or not to summarize the information per sample.

Value

A MoleculeExperiment object summary.

Examples

```
# get example data
repoDir <- system.file("extdata", package = "MoleculeExperiment")
repoDir <- paste0(repoDir, "/xenium_V1_FF_Mouse_Brain")
me <- readXenium(repoDir,
                 keepCols = "essential",
                 addBoundaries = "cell")

showMolecules(me)
```

```
showBoundaries(me)

nFeatures(me)
nFeatures(me, perSample = TRUE)

nTranscripts(me)
nTranscripts(me, perSample = TRUE)
```

Index