ArrayExpressHTS

March 24, 2012

`ArrayExpressHTS`	*ExpressionSet for RNA-Seq experiment submitted in ArrayExpress and ENA*

Description

`ArrayExpressHTS` runs the RNA-Seq pipeline on a transcription profiling experiment available on the ArrayExpress database and produces an `ExpressionSet` R object. `ArrayExpressHTS` requires an Internet connection.

Usage

```
ArrayExpressHTS(accession, usercloud = TRUE,
    options = getAEDefaultOptions(), nnodes = 10,
    pool = "32G", attempts = 4, dir = getwd(),
    refdir = getDefaultReferenceDir(), filter = TRUE,
    want.reports = TRUE )
```

Arguments

`accession`	an ArrayExpress experiment accession identifier.
`usercloud`	if TRUE, the R-Cloud will be used to schedule and process the experiment in parallel, otherwise data files are processed sequentially.
`options`	default options .
`nnodes`	if set, the selected amount of nodes will be allocated when R-Cloud cluster is created. Not used when `usercloud` is set to FALSE.
`pool`	server pool, from which cluster nodes are allocated. Allowed values are: 'default', '4G', '8G', '16G', '32G'. Not used when `usercloud` is set to FALSE.
`attempts`	number of attempts the package uses to allocate server node before giving up. Not used when `usercloud` is set to FALSE.
`dir`	folder where experiment data will be stored and processed. Default is current directory.
`refdir`	the directory where reference data is located.
`filter`	if TRUE, data filtering will be used as part of the the pipeline.
`want.reports`	if TRUE, quality reports are produced, however, it usually takes longer and more memory is used. For faster computation, set to FALSE.

Value

The output is an object of class ExpressionSet containing expression values in assayData (corresponding to the raw sequencing data files), the information contained in the .sdrf file in phenoData, the information in the adf file in featureData and the idf file content in experimentData.

If executed on a local PC, make sure that tools are available to the pipeline. Check prepareAnnotation to see what needs to be done to make tools available.

Author(s)

Andrew Tikhonov <andrew@ebi.ac.uk>, Angela Goncalves <angela.goncalves@ebi.ac.uk>

See Also

ArrayExpressHTSFastQ, prepareReference, prepareAnnotation, prepareAnnotation

Examples

```
if (isRCloud()) { # disabled on local configs so as not to kill package building process

    # if executed on a local PC, make sure tools
    # are available to the pipeline. Check initDefaultEnvironment
    # help page for instructions.
    expfolder = tempdir();

    # run the pipeline
    #
    aehts = ArrayExpressHTS("E-GEOD-16190", dir = expfolder);

    # load the expression set object
    loadednames = load(paste(expfolder, "/E-GEOD-16190/eset_notstd_rpkm.RData", sep=""));
    loadednames;

    get('library')(Biobase);

    # print out the expression values
    #
    head(assayData(eset)$exprs);

    # print out the experiment meta data
    experimentData(eset);
    pData(eset);
}
```

ArrayExpressHTSFastQ

ExpressionSet for RNA-Seq raw data files

Description

ArrayExpressHTSFastQ runs the RNA-Seq pipeline on raw RNA-Seq data files and an .sdrf experiment descriptor and produces an ExpressionSet R object.

Usage

```
ArrayExpressHTSFastQ(accession, organism, quality = c("auto","FastqQuality",
    "SFastqQuality"), usercloud = TRUE, options = getAEDefaultOptions(), nno
    pool = "32G", attempts = 4, dir = getwd(), refdir = getDefaultReferenceD
    filter = TRUE, want.reports = TRUE)
```

Arguments

accession	name of the folder where experiment data is stored. Experiment description files .sdrf and .idf should be stored in this folder. The actual data files should be stored in the 'data' that should be created in this folder.
organism	this parameter is used to select appropriate alignment reference.
quality	this parameter is used to explicitly select the quality scale used in the fastq data files. By default "auto" is used. In case quality scale cannot be automatically detected, the user will be prompted to increase the detection depth or set the quality scale manually. "FastqQuality" corresponds to Phred+33, "SFastqQuality" corresponds to Phred+64.
usercloud	if TRUE, the R-Cloud will be used to schedule and compute the experiment in parallel, otherwise the data files are computed sequentially.
options	default options .
nnodes	if set, the selected amount of nodes will be allocated when the R-Cloud cluster is created. Not used when usercloud is set to FALSE.
pool	server pool, from which cluster nodes are allocated. Allowed values are: 'default', '4G', '8G', '16G', '32G'. Not used when usercloud is set to FALSE.
attempts	number of attempts the package uses to allocate server node before giving up. Not used when usercloud is set to FALSE.
dir	folder where experiment data will be stored and processed. Default is current directory.
refdir	directory where the reference data is located.
filter	if TRUE, data filtering will be used as part of the the pipeline.
want.reports	if TRUE, quality reports are produced, however, it usually takes longer and more memory is used. For faster computation, set to FALSE.

Value

The output is an object of class ExpressionSet containing expression values in assayData (corresponding to the raw sequencing data files), the information contained in the .sdrf file in phenoData, the information in the adf file in featureData and the idf file content in experimentData.

Author(s)

Andrew Tikhonov Angela Goncalves Maintainer: <andrew@ebi.ac.uk> Maintainer: <angela.goncalves@ebi.ac.uk>

See Also

ArrayExpressHTS, prepareReference, prepareAnnotation

Examples

```
if (isRCloud()) { # disabled on local configs so as not to kill package building process

    # In ArrayExpressHTS/expdata there is testExperiment, which is
    # a very short version of E-GEOD-16190 experiment, placed there
    # for testing reasons.
    #
    # Experiment in ArrayExpress:
    # http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-16190
    #
    # the following piece of code will take ~1.5 hours to compute
    # on local PC and ~10 minutes on R-Cloud
    #

    # if executed on a local PC, make sure tools are available
    # to the pipeline. Check initDefaultEnvironment help page
    # for instructions.
    #

    # create a temporary folder where experiment will be copied
    # computing experiment in the package folder may cause issues
    # with file permissions and therefore failures.
    #
    #
    srcfolder <- system.file("expdata", "testExperiment", package="ArrayExpressHTS");

    dstfolder <- tempdir();

    file.copy(srcfolder, dstfolder, recursive = TRUE);

    # run the pipeline
    #
    # set usercloud = FALSE if executing on local PC,
    # therefore parallel computation will be disabled
    #
    aehts = ArrayExpressHTSFastQ(accession = "testExperiment",
        organism = "Homo_sapiens", dir = dstfolder);

    # load the expression set object
    loadednames = load(paste(dstfolder, "/testExperiment/eset_notstd_rpkm.RData", sep=""))
    loadednames;

    get('library')(Biobase);

    # print out the expression values
    #
    head(assayData(eset)$exprs);

    # print out the experiment meta data
    experimentData(eset);
    pData(eset);
}
```

getPipelineOption *Get ArrayExpressHTS option*

Description

`getPipelineOption` returns ArrayExpressHTS option specified by option name.

Usage

```
getPipelineOption(name, options = defaultOptions)
```

Arguments

`name`	name of the option to return. In order to see all options, use `getPipelineOptions`;
`options`	options environment, by default the ArrayExpresHTS option environment;

Value

The result is the value of the option.

Author(s)

Andrew Tikhonov <andrew@ebi.ac.uk>, Angela Goncalves <angela.goncalves@ebi.ac.uk>

See Also

`getPipelineOptions`, `setPipelineOptions`, `initDefaultEnvironment`, `ArrayExpressHTS`

Examples

```
getPipelineOption("trace");
getPipelineOption("cufflinks");
getPipelineOption("bowtie");
getPipelineOption("tophat");
getPipelineOption("samtools");
```

getPipelineOptions *Get ArrayExpressHTS options*

Description

`getPipelineOptions` returns an environment where ArrayExpressHTS options are stored.

Usage

```
getPipelineOptions()
```

Value

The output is the environment where pipeline options are stored.

Author(s)

Andrew Tikhonov <andrew@ebi.ac.uk>, Angela Goncalves <angela.goncalves@ebi.ac.uk>

See Also

getPipelineOption, setPipelineOptions, initDefaultEnvironment, ArrayExpressHTS

Examples

```
# get options
pipelineOptions = getPipelineOptions()

# list all names
ls(pipelineOptions)
```

---

initDefaultEnvironment

Initialize ArrayExpressHTS default environment

Description

initDefaultEnvironment initializes 'PATH' environment variable with paths to the tools used by ArrayExpressHTS. The function can be called to reinitialize the environment if any of the paths needs to be changed or if the package is executed in a local PC configuration.

Paths to pipeline tools are stored in the following pipeline options 'bowtie', 'tophat', 'bwa', 'cuflinks', 'mmseq', 'samtools', etc. If you need to change any of the paths, use setPipelineOptions. To check current values, use getPipelineOption. To see all options, use getPipelineOptions.

When options, namely the paths, have been changed call 'initDefaultEnvironment' to reinitialize environment. If the pipeline is executed on the R-Cloud, there is generally no necessity to setup anything since everything has been setup to a proper working configuration.

Plase note that some 'tophat' and 'cuflinks' versions are not fully compatible and cannot be automatically executed by the pipeline without intervension. Use the versions specified in the options, these versions are used in the R-Cloud. Other versions can be used at user"s own risk.

Usage

```
initDefaultEnvironment()
```

Arguments

No arguments.

Value

No output.

Author(s)

Andrew Tikhonov <andrew@ebi.ac.uk>, Angela Goncalves <angela.goncalves@ebi.ac.uk>

See Also

getPipelineOption, getPipelineOptions, setPipelineOptions, ArrayExpressHTS

Examples

```
initDefaultEnvironment();
```

---

isRCloud *Check the code is running on R-Cloud*

Description

isRCloud returns TRUE/FALSE indicating whether the configuration is an R-Cloud.

Usage

```
isRCloud()
```

Value

If TRUE, the configuration is the R-Cloud.

Author(s)

Andrew Tikhonov <andrew@ebi.ac.uk>, Angela Goncalves <angela.goncalves@ebi.ac.uk>

See Also

getPipelineOption, setPipelineOptions, initDefaultEnvironment, ArrayExpressHTS

Examples

```
if ( isRCloud() ) {

    # we're on the R-Cloud
    print("R-Cloud configuration");
} else {

    # we're somewhere else
    print("Other configuration");
}
```

`prepareAnnotation` *Prepare annotation data for the RNA-Seq Pipeline*

Description

`prepareAnnotation` downloads the required annotation file for the selected organism from Ensembl and processes it so that it can be used by the pipeline. `prepareAnnotation` requires an Internet connection.

Usage

```
prepareAnnotation(organism, version = "current",
    location = getDefaultReferenceDir(), refresh = FALSE, run = TRUE)
```

Arguments

`organism`	supported organism names can be viewed in the Ensemble database. Check 'ftp://ftp.ensembl.org/pub'.
`version`	'current' or other appropriate version. Check 'ftp://ftp.ensembl.org/pub'.
`location`	indicates where the annotation data should be stored.
`refresh`	if TRUE, existing annotation data will be rebuilt.
`run`	if FALSE, the commands to obtain and process the annotation will not be executed.

Value

The output is the version of the organism annotation that has been downloaded and processed. The annotation files are kept in the folder defined in `location` parameter.

Author(s)

Andrew Tikhonov <andrew@ebi.ac.uk>, Angela Goncalves <angela.goncalves@ebi.ac.uk>

See Also

`ArrayExpressHTS`, `ArrayExpressHTSFastQ`, `prepareReference`

Examples

```
if (isRCloud()) { # disabled on local configs so as not to kill package building process

    par(ask = FALSE)

    # the following piece of code will take ~1.5 hours to complete
    #

    # if executed on a local PC, make sure tools are available
    # to the pipeline. Check initDefaultEnvironment help page
    # for instructions.
    #
```

```
    # create directory
    #
    # Please note, tempdir() is used for automamtic test
    # execution. Select directory more appropriate and
    # suitable for keeping reference data.
    #
    referencefolder = paste(tempdir(), "/reference", sep = "")

    dir.create(referencefolder)

    # download and prepare annotation
    prepareAnnotation("Homo_sapiens", "current", location = referencefolder)
    prepareAnnotation("Mus_musculus", "NCBIM37.61", location = referencefolder)

}
```

`prepareReference` *Prepare reference data for the RNA-Seq Pipeline*

Description

`prepareReference` downloads reference genome or transcriptome for the selected organism from the Ensembl database and processes it so that it can be used by the pipeline. `prepareReference` requires an Internet connection.

Usage

```
prepareReference(organism, version = "current",
    type = c("genome", "transcriptome"),
    location = getDefaultReferenceDir(),
    aligner = c("bwa", "bowtie", "tophat"),
    refresh = FALSE, run = TRUE)
```

Arguments

`organism`	supported organism names can be viewed in the Ensemble database. Check 'ftp://ftp.ensembl.org/pub'.
`version`	'current' or other appropriate version. Check 'ftp://ftp.ensembl.org/pub'.
`type`	two values are supported: "genome", "transcriptome"
`location`	indicates where the reference data should be stored.
`aligner`	3 types of aligners are supported: 'bwa', 'bowtie' and 'tophat'.
`refresh`	if TRUE, existing reference data will be rebuilt.
`run`	if FALSE, the downloading and processing commands will not be executed.

Value

The output is the version of the organism reference that has been downloaded and processed. The reference files are kept in the folder defined in `location` parameter.

Author(s)

Andrew Tikhonov <andrew@ebi.ac.uk>, Angela Goncalves <angela.goncalves@ebi.ac.uk>

See Also

ArrayExpressHTS, ArrayExpressHTSFastQ, prepareAnnotation

Examples

```
if (isRCloud()) {

    par(ask = FALSE)

    # the following piece of code will take ~3 hours to complete
    #

    # if executed on a local PC, make sure tools are available
    # to the pipeline. Check initDefaultEnvironment help page
    # for instructions.
    #

    # create directory
    #
    # Please note, tempdir() is used for automamtic test
    # execution. Select directory more appropriate and
    # suitable for keeping reference data.
    #
    referencefolder = paste(tempdir(), "/reference", sep = "")

    dir.create(referencefolder)

    # download and prepare reference
    prepareReference("Homo_sapiens", version = "GRCh37.61",
        type = "genome", aligner = "bowtie", location = referencefolder )
    prepareReference("Homo_sapiens", version = "GRCh37.61",
        type = "transcriptome", aligner = "bowtie", location = referencefolder )
    prepareReference("Mus_musculus", version = "current",
        type = "genome", aligner = "bowtie", location = referencefolder )
    prepareReference("Mus_musculus", version = "current",
        type = "transcriptome", aligner = "bowtie", location = referencefolder )
}
```

---

setPipelineOptions *Set ArrayExpressHTS options*

Description

setPipelineOptions sets one or a number of ArrayExpressHTS options.

Usage

```
setPipelineOptions(...)
```

Arguments

...	a list ot options to set.

Value

No output.

Author(s)

Andrew Tikhonov <andrew@ebi.ac.uk>, Angela Goncalves <angela.goncalves@ebi.ac.uk>

See Also

getPipelineOption, getPipelineOptions, initDefaultEnvironment, ArrayExpressHTS

Examples

```
setPipelineOptions("trace" = "disabled", "memorymonitor" = "disabled");
```

Index