

Package ‘systemPipeR’

April 15, 2020

Type Package

Title systemPipeR: NGS workflow and report generation environment

Version 1.20.0

Date 2019-10-25

Author Thomas Girke

Maintainer Thomas Girke <thomas.girke@ucr.edu>

biocViews Genetics, Infrastructure, DataImport, Sequencing, RNASeq, RiboSeq, ChIPSeq, MethylSeq, SNP, GeneExpression, Coverage, GeneSetEnrichment, Alignment, QualityControl, ImmunoOncology

Description R package for building and running automated end-to-end analysis workflows for a wide range of next generation sequence (NGS) applications such as RNA-Seq, ChIP-Seq, VAR-Seq and Ribo-Seq. Important features include a uniform workflow interface across different NGS applications, automated report generation, and support for running both R and command-line software, such as NGS aligners or peak/variant callers, on local computers or compute clusters. Efficient handling of complex sample sets and experimental designs is facilitated by a consistently implemented sample annotation infrastructure. Instructions for using systemPipeR are given in the Overview Vignette (HTML). The remaining Vignettes, linked below, are workflow templates for common NGS use cases.

Depends Rsamtools (>= 1.31.2), Biostrings, ShortRead (>= 1.37.1), methods

Imports BiocGenerics, GenomicRanges, GenomicFeatures (>= 1.31.3), SummarizedExperiment, VariantAnnotation (>= 1.25.11), rjson, ggplot2, grid, limma, edgeR, DESeq2, GOstats, GO.db, annotate, pheatmap, batchtools, yaml

Suggests ape, RUnit, BiocStyle, knitr, rmarkdown, biomaRt, BiocParallel, BiocManager

VignetteBuilder knitr

SystemRequirements systemPipeR can be used to run external command-line software (e.g. short read aligners), but the corresponding tool needs to be installed on a system.

License Artistic-2.0

URL <http://girke.bioinformatics.ucr.edu/systemPipeR/>

git_url <https://git.bioconductor.org/packages/systemPipeR>

git_branch RELEASE_3_10

git_last_commit 70b884d

git_last_commit_date 2019-10-29

Date/Publication 2020-04-14

R topics documented:

alignStats	3
catDB-class	4
catmap	5
clusterRun	6
countRangeset	8
createWF	10
featureCoverage	11
featuretypeCounts	14
filterDEGs	16
filterVars	17
genFeatures	19
getQsubargs	21
GOHyperGAll	22
INTERSECTset-class	25
loadWorkflow	26
mergeBamByFactor	28
module	29
moduleload	30
olBarplot	31
olRanges	33
output_update	33
overLapper	35
plotfeatureCoverage	37
plotfeaturetypeCounts	39
predORF	41
preprocessReads	43
qsubRun	44
readComp	45
renderWF	46
returnRPKM	47
runCommandline	48
runDiff	50
run_DESeq2	51
run_edgeR	52
run_track	53
scaleRanges	54
seeFastq	55
subsetWF	57
symLink2bam	58
sysargs	59
SYSargs-class	59
SYSargs2-class	61
SYSargs2list	63
SYSargs2Pipe-class	64
SYSargs2Pipe_ls	66

<i>alignStats</i>	3
systemArgs	67
targets.as.df	68
variantReport	69
vennPlot	71
VENNset-class	74
writeTargetsout	75
writeTargetsRef	77
Index	78

<code>alignStats</code>	<i>Alignment statistics</i>
-------------------------	-----------------------------

Description

Generate data frame containing important read alignment statistics such as the total number of reads in the FASTQ files, the number of total alignments, as well as the number of primary alignments in the corresponding BAM files.

Usage

```
alignStats(args, output_index = 1)
```

Arguments

<code>args</code>	Object of class <code>SYSargs</code> or <code>SYSargs2</code> .
<code>output_index</code>	A numeric index positions of the file in <code>SYSargs2</code> object, slot <code>output</code> . Default is <code>output_index=1</code> .

Value

`data.frame` with alignment statistics.

Author(s)

Thomas Girke

See Also

`clusterRun` and `runCommandline` and `output_update`

Examples

```
#####
## Examples with \code{SYSargs} object ##
#####
## Construct SYSargs object from param and targets files
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
args
names(args); modules(args); cores(args); outpaths(args); sysargs(args)

## Not run:
```

```

## Execute SYSargs on single machine
runCommandline(args=args)

## Alignment stats
read_statsDF <- alignStats(args)
read_statsDF <- cbind(read_statsDF[targets$FileName,], targets)
write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE, sep="\t")

## End(Not run)

#####
## Examples with \code{SYSargs2} object ##
#####
## Construct SYSargs2 object from CWL param, CWL input, and targets files
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF
targets <- system.file("extdata", "targets.txt", package="systemPipeR")

names(WF); modules(WF); targets(WF)[1]; cmdlist(WF)[1:2]; output(WF)

## Not run:
## Execute SYSargs2 on single machine
WF <- runCommandline(args=WF)

## Alignment stats
read_statsDF <- alignStats(WF)
read_statsDF <- cbind(read_statsDF[targets$FileName,], targets)
write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE, sep="\t")

## End(Not run)

```

catDB-class

Class "catDB"

Description

Container for storing mappings of genes to annotation categories such as gene ontologies (GO), pathways or conserved sequence domains. The `catmap` slot stores a list of `data.frames` providing the direct assignments of genes to annotation categories (e.g. gene-to-GO mappings); `catlist` is a list of lists of all direct and indirect associations to the annotation categories (e.g. genes mapped to a pathway); and `idconv` allows to store a lookup-table for converting identifiers (e.g. array feature ids to gene ids).

Objects from the Class

Objects can be created by calls of the form `new("catDB", ...)`.

Slots

`catmap`: Object of class "list" list of `data.frames`

catlist: Object of class "list" list of lists
idconv: Object of class "ANY" list of data.frames

Methods

catlist signature(x = "catDB"): extracts data from catlist slot
catmap signature(x = "catDB"): extracts data from catmap slot
coerce signature(from = "list", to = "catDB"): as(list, "catDB")
idconv signature(x = "catDB"): extracts data from idconv slot
names signature(x = "catDB"): extracts slot names
show signature(object = "catDB"): summary view of catDB objects

Author(s)

Thomas Girke

See Also

makeCATdb, GOHyperGAll, GOHyperGAll_Subset, GOHyperGAll_Simplify, GOCluster_Report, goBarplot

Examples

```
showClass("catDB")
## Not run:
## Obtain annotations from BioMart
listMarts() # To choose BioMart database
m <- useMart("ENSEMBL_MART_PLANT"); listDatasets(m)
m <- useMart("ENSEMBL_MART_PLANT", dataset="athaliana_eg_gene")
listAttributes(m) # Choose data types you want to download
go <- getBM(attributes=c("go_accession", "tair_locus", "go_namespace_1003"), mart=m)
go <- go[go[,3]!="",,]; go[,3] <- as.character(go[,3])
write.table(go, "GOannotationsBiomart_mod.txt", quote=FALSE, row.names=FALSE, col.names=FALSE, sep="\t")

## Create catDB instance (takes a while but needs to be done only once)
catdb <- makeCATdb(myfile="GOannotationsBiomart_mod.txt", lib=NULL, org="", colno=c(1,2,3), idconv=NULL)
catdb

## End(Not run)
```

catmap

catDB accessor methods

Description

Methods to access information from catDB object.

Usage

catmap(x)

Arguments

x object of class catDB

Value

various outputs

Author(s)

Thomas Girke

Examples

```
## Not run:
## Obtain annotations from BioMart
m <- useMart("ENSEMBL_MART_PLANT"); listDatasets(m)
m <- useMart("ENSEMBL_MART_PLANT", dataset="athaliana_eg_gene")
listAttributes(m) # Choose data types you want to download
go <- getBM(attributes=c("go_accession", "tair_locus", "go_namespace_1003"), mart=m)
go <- go[go[,3]!="",,]; go[,3] <- as.character(go[,3])
write.table(go, "GOannotationsBiomart_mod.txt", quote=FALSE, row.names=FALSE, col.names=FALSE, sep="\t")

## Create catDB instance (takes a while but needs to be done only once)
catdb <- makeCATdb(myfile="GOannotationsBiomart_mod.txt", lib=NULL, org="", colno=c(1,2,3), idconv=NULL)
catdb

## Access methods for catDB
catmap(catdb)$D_MF[1:4,]
catlist(catdb)$L_MF[1:4]
idconv(catdb)

## End(Not run)
```

clusterRun

Submit command-line tools to cluster

Description

Submits non-R command-line software to queueing/scheduling systems of compute clusters using run specifications defined by functions similar to runCommandline. clusterRun can be used with most queueing systems since it is based on utilities from the batchtools package which supports the use of template files (*.tmpl) for defining the run parameters of the different schedulers. The path to the *.tmpl file needs to be specified in a conf file provided under the conffile argument.

Usage

```
clusterRun(args, FUN = runCommandline, more.args = list(args = args, make_bam = TRUE), conffile = ".b
```

Arguments

args	Object of class SYSargs or SYSargs2.
FUN	Accepts functions such as runCommandLine(args, ...) where the args argument is mandatory and needs to be of class SYSargs or SYSargs2.
more.args	Object of class list, which provides the arguments that control the FUN function.
confFile	Path to conf file (default location <code>./.batchtools.conf.R</code>). This file contains in its simplest form just one command, such as this line for the Slurm scheduler: <code>cluster.functions <-makeClusterFunctionsSlurm(template="batchtools.slurm.tpl")</code> . For more detailed information visit this page: https://mllg.github.io/batchtools/index.html
template	The template files for a specific queueing/scheduling systems can be downloaded from here: https://github.com/mllg/batchtools/tree/master/inst/templates . Slurm, PBS/Torque, and Sun Grid Engine (SGE) templates are provided.
Njobs	Integer defining the number of cluster jobs. For instance, if args contains 18 command-line jobs and Njobs=9, then the function will distribute them across 9 cluster jobs each running 2 command-line jobs. To increase the number of CPU cores used by each process, one can do this under the corresponding argument of the command-line tool, e.g. <code>-p</code> argument for Tophat.
runid	Run identifier used for log file to track system call commands. Default is "01".
resourceList	List for reserving for each cluster job sufficient computing resources including memory (Megabyte), number of nodes, CPU cores, walltime (minutes), etc. For more details, one can consult the template file for each queueing/scheduling system.

Value

Object of class Registry, as well as files and directories created by the executed command-line tools.

Author(s)

Daniela Cassol and Thomas Girke

References

For more details on batchtools, please consult the following page: <https://github.com/mllg/batchtools/>

See Also

clusterRun replaces the older functions getQsubargs and qsubRun.

Examples

```
#####
## Examples with \code{SYSargs} object ##
#####
## Construct SYSargs object from param and targets files
param <- system.file("extdata", "hisat2.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
args
names(args); modules(args); cores(args); outpaths(args); sysargs(args)
```

```

## Not run:
## Execute SYSargs on multiple machines of a compute cluster. The following
## example uses the conf and template files for the Slurm scheduler. Please
## read the instructions on how to obtain the corresponding files for other schedulers.
file.copy(system.file("extdata", ".batchtools.conf.R", package="systemPipeR"), ".")
file.copy(system.file("extdata", "batchtools.slurm.tmpl", package="systemPipeR"), ".")
resources <- list(walltime=120, ntasks=1, ncpus=cores(args), memory=1024)
reg <- clusterRun(args, FUN = runCommandline, more.args = list(args = args, make_bam = TRUE), conffile=".batcht

## Monitor progress of submitted jobs
getStatus(reg=reg)
file.exists(outpaths(args))

## End(Not run)

#####
## Examples with \code{SYSargs2} object ##
#####
## Construct SYSargs2 object from CWL param, CWL input, and targets files
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF
names(WF); modules(WF); targets(WF)[1]; cmdlist(WF)[1:2]; output(WF)

## Not run:
## Execute SYSargs2 on multiple machines of a compute cluster. The following
## example uses the conf and template files for the Slurm scheduler. Please
## read the instructions on how to obtain the corresponding files for other schedulers.
file.copy(system.file("extdata", ".batchtools.conf.R", package="systemPipeR"), ".")
file.copy(system.file("extdata", "batchtools.slurm.tmpl", package="systemPipeR"), ".")
resources <- list(walltime=120, ntasks=1, ncpus=4, memory=1024)
reg <- clusterRun(WF, FUN = runCommandline, more.args = list(args = WF, make_bam = TRUE), conffile=".batchtools

## Monitor progress of submitted jobs
getStatus(reg=reg)

## Updates the path in the object \code{output(WF)}
WF <- output_update(WF, dir=FALSE, replace=TRUE, extension=c(".sam", ".bam"))

## Alignment stats
read_statsDF <- alignStats(WF)
read_statsDF <- cbind(read_statsDF[targets$FileName,], targets)
write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE, sep="\t")

## End(Not run)

```


Description

Convenience function to perform read counting iteratively for several range sets, e.g. peak range sets or feature types. Internally, the read counting is performed with the `summarizeOverlaps` function from the `GenomicAlignments` package. The resulting count tables are directly saved to files.

Usage

```
countRangeset(bfl, args, format="tabular", ...)
```

Arguments

<code>bfl</code>	BamFileList object containing paths to one or more BAM files.
<code>args</code>	Object of class <code>SYSargs</code> or <code>SYSargs2</code> where <code>infile1(args)</code> specifies the paths to the tabular range data files (e.g. peak ranges) used for counting.
<code>format</code>	Format of input range files. Currently, supported are <code>tabular</code> or <code>bed</code> . If <code>tabular</code> is selected then the input range files need to contain the proper column titles to coerce with <code>as(..., "GRanges")</code> to <code>GRanges</code> objects after importing them with <code>read.delim</code> . The latter is the case for the peak files (<code>*peaks.xls</code>) generated by the MACS2 software.
<code>...</code>	Arguments to be passed on to internally used <code>summarizeOverlaps</code> function.

Value

Named character vector containing the paths from `outpaths(args)` to the resulting count table files.

Author(s)

Thomas Girke

See Also

`summarizeOverlaps`

Examples

```
## Paths to BAM files
param <- system.file("extdata", "bowtieSE.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args_bam <- systemArgs(sysma=param, mytargets=targets)
bfl <- BamFileList(outpaths(args_bam), yieldSize=50000, index=character())

## Not run:
#####
## Examples with \code{SYSargs} object ##
#####
## Construct SYSargs object
## SYSargs with paths to range data and count files
args <- systemArgs(sysma="param/count_rangesets.param", mytargets="targets_mac2.txt")

## Iterative read counting
countDFnames <- countRangeset(bfl, args, mode="Union", ignore.strand=TRUE)
writeTargetsout(x=args, file="targets_countDF.txt", overwrite=TRUE)
```

```
#####
## Examples with \code{SYSargs2} object ##
#####
## Construct SYSargs2 object
## SYSargs2 with paths to range data and count files
dir_path <- system.file("extdata/cwl/count_rangesets", package="systemPipeR")
args <- loadWF(targets = "targets_macs.txt", wf_file = "count_rangesets.cwl",
  input_file = "count_rangesets.yml", dir_path = dir_path)
args <- renderWF(args, inputvars = c(FileName = "_FASTQ_PATH1_", SampleName = "_SampleName_"))

## Iterative read counting
countDFnames <- countRangeset(bf1, args, mode="Union", ignore.strand=TRUE)

## End(Not run)
```

createWF

Constructs SYSargs2 object and creates CWL param files

Description

The constructor function creates an SYSargs2 S4 class object from three input files: a CWL param and input files, and one simple tabular files, a targets file. The latter is optional for workflow steps lacking input files. Also, the function creates and saves the CWL param files. The latest storages all the parameters required for running command-line software, following the standard and specification defined on [Common Workflow Language \(CWL\)](#).

Usage

```
createWF(targets = NULL, commandLine, results_path = "./results", module_load = "baseCommand", file
```

Arguments

targets	Path to targets file. Assign NULL to run the pipeline without 'targets' file. This can be useful for running specific workflows which do not require input files.
commandLine	Object of class list, which provides all the parameters required for running command-line software.
results_path	Path to the results folder.
module_load	Name of software to load as character. Default is "default", which creates a subfolder and two files: *.cwl and *.yml at ./param/cwl/.
file	Default is "default", which creates a subfolder and two files: *.cwl and *.yml at ./param/cwl/. If a different location and names are required, the names of the files can be specified as a character vector.
overwrite	If set to TRUE, existing files of the same name will be overwritten.
cwlVersion	version of the Common Workflow Language. More information here: https://www.commonwl.org/ .
class	Names of Common Workflow Language Specification. The following switches are accepted: CommandLineTool and Workflow.

Value

SYSargs2 object

Author(s)

Daniela Cassol and Thomas Girke

ReferencesFor more details on CWL, please consult the following page: <https://www.commonwl.org/>**See Also**

loadWorkflow renderWF showClass("SYSargs2")

Examples

```
# "hisat2 -S ./results/_SampleName_.sam -x ./data/tair10.fasta -k 1 --min-intronlen 30 --max-intronlen 3000
## Provide a list with all the arguments
baseCommand <- list(baseCommand="hisat2")
inputs <- list(
  "S"=list(type="File", preF="-S", yml="./results/_SampleName_.sam"),
  "x"=list(type="File", preF="-x", yml="./data/tair10.fasta"),
  "k"= list(type="int", preF="-k", yml= 1L),
  "threads"= list(type="int", preF="--threads", yml=4L),
  "min-intronlen"= list(type="int", preF="--min-intronlen", yml= 30L),
  "max-intronlen"= list(type="int", preF="--max-intronlen", yml=3000L),
  "U"=list(type="File", preF="-U", yml="./data/_FASTQ_PATH1_") )
outputs <- list("hisat2_sam"=list(type="File", ". /results/_SampleName_.sam"))
commandLine <- list(baseCommand=baseCommand, inputs=inputs, outputs=outputs)
## Not run:
## Create a SYSargs2 object and populate all the command-line
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
WF <- createWF(targets=targets, commandLine, results_path="./results", module_load="baseCommand", file = "def
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))

## End(Not run)
```

featureCoverage

*Genome read coverage by transcript models***Description**

Computes read coverage along single and multi component features based on genomic alignments. The coverage segments of component features are spliced to continuous ranges, such as exons to transcripts or CDSs to ORFs. The results can be obtained with single nucleotide resolution (e.g. around start and stop codons) or as mean coverage of relative bin sizes, such as 100 bins for each feature. The latter allows comparisons of coverage trends among transcripts of variable length. The results can be obtained for single or many features (e.g. any number of transcripts) at once. Visualization of the coverage results is facilitated by a downstream plotfeatureCoverage function.

Usage

```
featureCoverage(bfl, grl, resizereads = NULL, readlengthrange = NULL, Nbins = 20,
  method = mean, fixedmatrix, resizefeatures, upstream, downstream,
  outfile, overwrite = FALSE)
```

Arguments

<code>bf1</code>	Paths to BAM files provided as <code>BamFileList</code> object. The name slot of the BAM files will be used for naming samples in the results.
<code>gr1</code>	Genomic ranges provided as <code>GRangesList</code> typically generated from <code>txdb</code> instances with operations like: <code>cdsBy(txdb,"tx")</code> or <code>exonsBy(txdb,"tx")</code> . Single component features will be processed the same way as multi component features.
<code>resizereads</code>	Positive integer defining the length alignments should be resized to prior to the coverage calculation. <code>NULL</code> will omit the resizing step.
<code>readlengthrange</code>	Positive integer of length 2 determining the read length range to use for the coverage calculation. Reads falling outside of the specified length range will be excluded from the coverage calculation. For instance, <code>readlengthrange=c(30:40)</code> will base the coverage calculation on reads between 30 to 40 bps. Assigning <code>NULL</code> will skip this filtering step.
<code>Nbins</code>	Single positive integer defining the number of segments the coverage of each feature should be binned into in order to obtain coverage summaries of constant length, e.g. for plotting purposes.
<code>method</code>	Defines the summary statistics to use for binning. The default is <code>method=mean</code> .
<code>fixedmatrix</code>	If set to <code>TRUE</code> , a coverage matrix with single nucleotide resolution will be returned for any number of transcripts centered around precise anchor points in a genome annotation, such as stop/start codons or transcription start sites. For instance, a matrix with coverage information 20bps upstream and downstream of the stop/start codons can be obtained with <code>fixedmatrix=TRUE,upstream=20,downstream=20</code> along with a <code>gr1</code> instance containing the CDS exon ranges required for this operation, e.g. generated with <code>cdsBy(txdb,"tx")</code> .
<code>resizefeatures</code>	Needs to be set to <code>TRUE</code> when <code>fixedmatrix=TRUE</code> . Internally, this will use the <code>systemPipeR::resizeFeature</code> function to extend single and multi component features at their most left and most right end coordinates. The corresponding extension values are specified under the <code>upstream</code> and <code>downstream</code> arguments.
<code>upstream</code>	Single positive integer specifying the upstream extension length relative to the orientation of each feature in the genome. More details are given above.
<code>downstream</code>	Single positive integer specifying the downstream extension length relative to the orientation of each feature in the genome. More details are given above.
<code>outfile</code>	Default <code>NULL</code> omits writing of the results to a file. If a file name is specified then the results are written to a tabular file. If <code>bf1</code> contains the paths to several BAM files then the results will be appended to the same file where the first column specifies the sample labels. Redirecting the results to file is particularly useful when processing large files of many sample where computation times can be significant.
<code>overwrite</code>	If set to <code>TRUE</code> any existing file assigned to <code>outfile</code> will be overwritten.

Value

The function allows to return the following four distinct outputs. The settings to return these instances are illustrated below in the example section.

- (A) `data.frame` containing binned coverage where rows are features and columns coverage bins. The first four columns contain (i) the sample names, (ii) the number of total aligned reads in the corresponding BAM files (useful for normalization), (iii) the feature IDs, (iv) strand of the coverage. All following columns are numeric and contain the actual coverage data for the sense and antisense strand of each feature.
- (B) `data.frame` containing coverage with single nucleotide resolution around anchor points such as start and stop codons. The two matrix components are appended column-wise. To clearly distinguish the two data components, they are separated by a specialty column containing pipe characters. The first four columns are the same as described under (A). The column title for the anchor point is 0. For instance, if the features are CDSs then the first 0 corresponds to the first nucleotide of the start codon and the second 0 to the last nucleotide of the stop codon. Upstream and downstream positions are indicated by negative and positive column numbers, respectively.
- (C) `data.frame` containing combined results of (A) and (B) where the first set of columns contains to the coverage around the start codons, the second one the binned coverage of the CDSs and the third one the coverage around the stop codons separated by the same pipe columns mentioned under (B).
- (D) `Rle` list containing the nucleotide level coverage of each feature

Author(s)

Thomas Girke

See Also

`plotfeatureCoverage`

Examples

```
## Construct SYSargs object from param and targets files
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)

## Not run:
## Features from sample data of systemPipeRdata package
library(GenomicFeatures)
file <- system.file("extdata/annotation", "tair10.gff", package="systemPipeRdata")
txdb <- makeTxDbFromGFF(file=file, format="gff3", organism="Arabidopsis")

## (A) Generate binned coverage for two BAM files and 4 transcripts
grl <- cdsBy(txdb, "tx", use.names=TRUE)
fcov <- featureCoverage(bfl=BamFileList(outpaths(args)[1:2]), grl=grl[1:4], resizereads=NULL,
  readlengthrange=NULL, Nbins=20, method=mean, fixedmatrix=FALSE,
  resizefeatures=TRUE, upstream=20, downstream=20)
plotfeatureCoverage(covMA=fcov, method=mean, scales="fixed", scale_count_val=10^6)

## (B) Coverage matrix upstream and downstream of start/stop codons
fcov <- featureCoverage(bfl=BamFileList(outpaths(args)[1:2]), grl=grl[1:4], resizereads=NULL,
  readlengthrange=NULL, Nbins=NULL, method=mean, fixedmatrix=TRUE,
  resizefeatures=TRUE, upstream=20, downstream=20)
plotfeatureCoverage(covMA=fcov, method=mean, scales="fixed", scale_count_val=10^6)
```

```

## (C) Combined matrix for both binned and start/stop codon
fcov <- featureCoverage(bfl=BamFileList(outpaths(args)[1:2]), grl=grl[1:4], resizereads=NULL,
  readlengthrange=NULL, Nbins=20, method=mean, fixedmatrix=TRUE,
  resizefeatures=TRUE, upstream=20, downstream=20)
plotfeatureCoverage(covMA=fcov, method=mean, scales="fixed", scale_count_val=10^6)

## (D) Rle coverage objects one for each query feature
fcov <- featureCoverage(bfl=BamFileList(outpaths(args)[1:2]), grl=grl[1:4], resizereads=NULL,
  readlengthrange=NULL, Nbins=NULL, method=mean, fixedmatrix=FALSE,
  resizefeatures=TRUE, upstream=20, downstream=20)

## End(Not run)

```

featuretypeCounts *Plot read distribution across genomic features*

Description

Counts how many reads in short read alignment files (BAM format) overlap with entire annotation categories. This utility is useful for analyzing the distribution of the read mappings across feature types, e.g. coding versus non-coding genes. By default the read counts are reported for the sense and antisense strand of each feature type separately. To minimize memory consumption, the BAM files are processed in a stream using utilities from the Rsamtools and GenomicAlignment packages. The counts can be reported for each read length separately or as a single value for reads of any length. Subsequently, the counting results can be plotted with the associated plotfeaturetypeCounts function.

Usage

```
featuretypeCounts(bfl, grl, singleEnd = TRUE, readlength = NULL, type = "data.frame")
```

Arguments

bfl	BamFileList object containing the paths to the target BAM files stored on disk. Note, memory-efficient processing is achieved by streaming through BAM files rather than reading entire files into memory at once. The maximum number of alignments to process in each iteration is determined by the yieldSize value passed on to the BamFileList function. For details see ?BamFileList.
grl	GRangesList object containing in each list component the range set of a feature type. Typically, this object is generated with the genFeatures function. For details see the example section below and ?genFeatures.
singleEnd	Specifies whether the targets BAM files contain alignments for single-end (SE) or paired-end read data. TRUE is for SE and FALSE for PE data.
readlength	Integer vector specifying the read length values for which to report counts separately. If readlength=NULL the length of the reads will be ignored resulting in a single value for each feature type and strand. Note, for PE data the two reads in a pair may differ in length. In those cases the length of the two reads is averaged and then assigned to the corresponding length category after rounding the mean length to the closest integer. This is not an ideal solution but a reasonable compromise for the purpose of the summary statistics generated by featuretypeCounts.

`type` Determines whether the results are returned as `data.frame` (`type="data.frame"`) or as `list` (`type="list"`). Each list component contains the counting results for one BAM file and is named after the corresponding sample. The `data.frame` result contains this sample assignment information in a separate column.

Value

The results are returned as `data.frame` or `list` of `data.frames`. For details see above under `types` argument. The result `data.frames` contain the following columns in the given order:

<code>SampleName</code>	Sample names obtained from <code>BamFileList</code> object.
<code>Strand</code>	Sense or antisense strand of read mappings.
<code>Featuretype</code>	Name of feature type provided by <code>GRangesList</code> object. Note, the total number of aligned reads is reported under the special feature type <code>'N_total_aligned'</code> . This value is useful for scaling/normalization purposes in plots, e.g. counts per million reads.
<code>Featuretypelength</code>	Total genomic length of each reduced feature type in bases. This value is useful to normalize the read counts by genomic length units, e.g. in plots.
Subsequent columns	Counts for reads of any length or for individual read lengths.

Author(s)

Thomas Girke

See Also

`plotfeaturetypeCounts`, `genFeatures`

Examples

```
## Construct SYSargs object from param and targets files
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)

## Not run:
## Features from sample data of systemPipeRdata package
library(GenomicFeatures)
file <- system.file("extdata/annotation", "tair10.gff", package="systemPipeRdata")
txdb <- makeTxDbFromGFF(file=file, format="gff3", organism="Arabidopsis")
feat <- genFeatures(txdb, featuretype="all", reduce_ranges=TRUE, upstream=1000, downstream=0, verbose=TRUE)

## Generate and plot feature counts for specific read lengths
fc <- featuretypeCounts(bfl=BamFileList(outpaths(args), yieldSize=50000), grl=feat, singleEnd=TRUE, readlength=100)
p <- plotfeaturetypeCounts(x=fc, graphicsfile="featureCounts.pdf", graphicsformat="pdf", scales="fixed", any)

## Generate and plot feature counts for any read length
fc2 <- featuretypeCounts(bfl=BamFileList(outpaths(args), yieldSize=50000), grl=feat, singleEnd=TRUE, readlength=any)
p2 <- plotfeaturetypeCounts(x=featureCounts2, graphicsfile="featureCounts2.pdf", graphicsformat="pdf", scales="fixed", any)

## End(Not run)
```

filterDEGs

*Filter and plot DEG results***Description**

Filters and plots DEG results for a given set of sample comparisons. The gene identifiers of all (i) Up_or_Down, (ii) Up and (iii) Down regulated genes are stored as separate list components, while the corresponding summary statistics, stored in a fourth list component, is plotted in form of a stacked bar plot.

Usage

```
filterDEGs(degDF, filter, plot = TRUE)
```

Arguments

degDF	data.frame generated by run_edgeR
filter	Named vector with filter cutoffs of format <code>c(Fold=2, FDR=1)</code> where Fold refers to the fold change cutoff (unlogged) and FDR to the p-value cutoff.
plot	Allows to turn plotting behavior on and off with default set to TRUE.

Details

Currently, there is no community standard available how to calculate fold changes (here logFC) of genomic ranges, such as gene or feature ranges, to unambiguously refer to them as features with increased or decreased read abundance; or in case of gene expression experiments to up or down regulated genes, respectively. To be consistent within systemPipeR, the corresponding functions, such as filterDEGs, use here the following definition. Genomic ranges with positive logFC values are classified as up and those with negative logFC values as down. This means if a comparison among two samples a and b is specified in the corresponding targets file as a-b then the feature with a positive logFC has a higher `_normalized_` read count value in sample a than in sample b, and vice versa. To inverse this assignment, users want to change the specification of their chosen sample comparison(s) in the targets file accordingly, e.g. change a-b to b-a. Alternatively, one can swap the column order of the matrix assigned to the `cmp` argument of the `run_edgeR` or `run_DESeq2` functions. Users should also be aware that for logFC values close to zero (noise range), the direction of the fold change (sign of logFC) can be very sensitive to minor differences in the normalization method, while this assignment is much more robust for more pronounced changes or higher absolute logFC values.

Value

Returns list with four components

UporDown	List of up or down regulated gene/transcript identifiers meeting the chosen filter settings for all comparisons defined in data frames <code>pval</code> and <code>log2FC</code> .
Up	Same as above but only for up regulated genes/transcript.
Down	Same as above but only for down regulated genes/transcript.

Author(s)

Thomas Girke

See Also

run_edgeR

Examples

```
targetspath <- system.file("extdata", "targets.txt", package="systemPipeR")
targets <- read.delim(targetspath, comment="#")
cmp <- readComp(file=targetspath, format="matrix", delim="-")
countfile <- system.file("extdata", "countDFeByg.xls", package="systemPipeR")
countDF <- read.delim(countfile, row.names=1)
edgeDF <- run_edgeR(countDF=countDF, targets=targets, cmp=cmp[[1]], independent=FALSE, mdsplot="")
pval <- edgeDF[, grep("_FDR$", colnames(edgeDF)), drop=FALSE]
fold <- edgeDF[, grep("_logFC$", colnames(edgeDF)), drop=FALSE]
DEG_list <- filterDEGs(degDF=edgeDF, filter=c(Fold=2, FDR=10))
names(DEG_list)
DEG_list$Summary
```

filterVars

Filter VCF files

Description

Convenience function for filtering VCF files based on user definable quality parameters. The function imports each VCF file into R, applies the filtering on an internally generated VRanges object and then writes the results to a new VCF file.

Usage

```
filterVars(args, filter, varcaller, organism)
```

Arguments

args	Object of class <code>SYSargs</code> or <code>SYSargs2</code> . The paths of the input VCF files are specified under <code>infile1(args)</code> and the paths of the output files under <code>outfile1(args)</code> or <code>output(args)</code> .
filter	Character vector of length one specifying the filter syntax that will be applied to the internally created <code>VRanges</code> object.
varcaller	Character vector of length one specifying the variant caller used for generating the input VCFs. Currently, this argument can be assigned 'gatk', 'bcftools' or 'vartools'.
organism	Character vector specifying the organism name of the reference genome.

Value

Output files in VCF format. Their paths can be obtained with `outpaths(args)` or `output(args)`.

Author(s)

Thomas Girke

See Also

variantReport combineVarReports, varSummar

Examples

```
## Alignment with BWA (sequentially on single machine)
param <- system.file("extdata", "bwa.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
sysargs(args)[1]

## Not run:
system("bwa index -a bwtsv ./data/tair10.fasta")
bampaths <- runCommandline(args=args)

## Alignment with BWA (parallelized on compute cluster)
resources <- list(walltime="20:00:00", nodes=paste0("1:ppn=", cores(args)), memory="10gb")
reg <- clusterRun(args, conffile=".BatchJobs.R", template="torque.tmpl", Njobs=18, runid="01",
                  resourceList=resources)

## Variant calling with GATK
## The following creates in the initial step a new targets file
## (targets_bam.txt). The first column of this file gives the paths to
## the BAM files created in the alignment step. The new targets file and the
## parameter file gatk.param are used to create a new SYSargs
## instance for running GATK. Since GATK involves many processing steps, it is
## executed by a bash script gatk_run.sh where the user can specify the
## detailed run parameters. All three files are expected to be located in the
## current working directory. Samples files for gatk.param and
## gatk_run.sh are available in the subdirectory ./inst/extdata/ of the
## source file of the systemPipeR package.
writeTargetsout(x=args, file="targets_bam.txt")
system("java -jar CreateSequenceDictionary.jar R=./data/tair10.fasta O=./data/tair10.dict")
# system("java -jar /opt/picard/1.81/CreateSequenceDictionary.jar R=./data/tair10.fasta O=./data/tair10.dict")
args <- systemArgs(sysma="gatk.param", mytargets="targets_bam.txt")
resources <- list(walltime="20:00:00", nodes=paste0("1:ppn=", 1), memory="10gb")
reg <- clusterRun(args, conffile=".BatchJobs.R", template="torque.tmpl", Njobs=18, runid="01",
                  resourceList=resources)
writeTargetsout(x=args, file="targets_gatk.txt")

## Variant calling with BCFtools
## The following runs the variant calling with BCFtools. This step requires in
## the current working directory the parameter file sambcf.param and the
## bash script sambcf_run.sh.
args <- systemArgs(sysma="sambcf.param", mytargets="targets_bam.txt")
resources <- list(walltime="20:00:00", nodes=paste0("1:ppn=", 1), memory="10gb")
reg <- clusterRun(args, conffile=".BatchJobs.R", template="torque.tmpl", Njobs=18, runid="01",
                  resourceList=resources)
writeTargetsout(x=args, file="targets_sambcf.txt")

## Filtering of VCF files generated by GATK
args <- systemArgs(sysma="filter_gatk.param", mytargets="targets_gatk.txt")
filter <- "totalDepth(vr) >= 2 & (altDepth(vr) / totalDepth(vr) >= 0.8) & rowSums(softFilterMatrix(vr))==4"
# filter <- "totalDepth(vr) >= 20 & (altDepth(vr) / totalDepth(vr) >= 0.8) & rowSums(softFilterMatrix(vr))==6"
filterVars(args, filter, varcaller="gatk", organism="A. thaliana")
writeTargetsout(x=args, file="targets_gatk_filtered.txt")
```

```

## Filtering of VCF files generated by BCFtools
args <- systemArgs(sysma="filter_sambcf.param", mytargets="targets_sambcf.txt")
filter <- "rowSums(vr) >= 2 & (rowSums(vr[,3:4])/rowSums(vr[,1:4]) >= 0.8)"
# filter <- "rowSums(vr) >= 20 & (rowSums(vr[,3:4])/rowSums(vr[,1:4]) >= 0.8)"
filterVars(args, filter, varcaller="bcftools", organism="A. thaliana")
writeTargetsout(x=args, file="targets_sambcf_filtered.txt")

## Annotate filtered variants from GATK
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_gatk_filtered.txt")
txdb <- loadDb("./data/tair10.sqlite")
fa <- FaFile(systemPipeR::reference(args))
variantReport(args=args, txdb=txdb, fa=fa, organism="A. thaliana")

## Annotate filtered variants from BCFtools
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
txdb <- loadDb("./data/tair10.sqlite")
fa <- FaFile(systemPipeR::reference(args))
variantReport(args=args, txdb=txdb, fa=fa, organism="A. thaliana")

## Combine results from GATK
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_gatk_filtered.txt")
combinedDF <- combineVarReports(args, filtercol=c(Consequence="nonsynonymous"))
write.table(combinedDF, "./results/combinedDF_nonsyn_gatk.xls", quote=FALSE, row.names=FALSE, sep="\t")

## Combine results from BCFtools
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
combinedDF <- combineVarReports(args, filtercol=c(Consequence="nonsynonymous"))
write.table(combinedDF, "./results/combinedDF_nonsyn_sambcf.xls", quote=FALSE, row.names=FALSE, sep="\t")

## Summary for GATK
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_gatk_filtered.txt")
write.table(varSummary(args), "./results/variantStats_gatk.xls", quote=FALSE, col.names = NA, sep="\t")

## Summary for BCFtools
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
write.table(varSummary(args), "./results/variantStats_sambcf.xls", quote=FALSE, col.names = NA, sep="\t")

## Venn diagram of variants
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_gatk_filtered.txt")
varlist <- sapply(names(outpaths(args))[1:4], function(x) as.character(read.delim(outpaths(args)[x])$VARID))
vennset_gatk <- overLapper(varlist, type="vennsets")
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
varlist <- sapply(names(outpaths(args))[1:4], function(x) as.character(read.delim(outpaths(args)[x])$VARID))
vennset_bcf <- overLapper(varlist, type="vennsets")
vennPlot(list(vennset_gatk, vennset_bcf), mymain="", mysub="GATK: red; BCFtools: blue", colmode=2, ccol=c("bl

## End(Not run)

```

genFeatures

Generate feature ranges from TxDb

Description

Function to generate a variety of feature types from TxDb objects using utilities provided by the GenomicFeatures package. The feature types are organized per gene and can be returned on that

level in their non-reduced or reduced form.

Currently, supported features include intergenic, promoter, intron, exon, cds, 5'/3'UTR and different transcript types. The latter contains as many transcript types as available in the tx_type column when extracting transcripts from TxDb objects as follows: transcripts(txdb,c("tx_name", "gene_id", "tx_t

Usage

```
genFeatures(txdb, featuretype = "all", reduce_ranges, upstream = 1000, downstream = 0, verbose = TRUE)
```

Arguments

txdb	TxDb object
featuretype	Feature types can be specified by assigning a character vector containing any of the following: c("tx_type", "promoter", "intron", "exon", "cds", "fiveUTR", "threeUTR", "intergenic"). The default all is a shorthand to select all supported features.
reduce_ranges	If set to TRUE the feature ranges will be reduced on the gene level. As a result overlapping feature components of the same type and from the same gene will be merged to a single range, e.g. two overlapping exons from the same gene are merged to one. Intergenic ranges are not affected by this setting. Note, all reduced feature types are labeled with the suffix '_red'.
upstream	Defines for promoter features the number of bases upstream from the transcription start site.
downstream	Defines for promoter features the number of bases downstream from the transcription start site.
verbose	verbose=FALSE turns off all print messages.

Value

The results are returned as a GRangesList where each component is a GRanges object containing the range set of each feature type. Intergenic ranges are assigned unique identifiers and recorded in the featuretype_id column of the metadata block. For this the ids of their adjacent genes are concatenated with two underscores as separator. If the adjacent genes overlap with other genes then their identifiers are included in the id string as well and separated by a single underscore.

Author(s)

Thomas Girke

See Also

transcripts and associated TxDb accessor functions from the GenomicFeatures package.

Examples

```
## Sample from GenomicFeatures package
library(GenomicFeatures)
gffFile <- system.file("extdata", "GFF3_files", "a.gff3", package="GenomicFeatures")
txdb <- makeTxDbFromGFF(file=gffFile, format="gff3", organism="Solanum lycopersicum")
feat <- genFeatures(txdb, featuretype="all", reduce_ranges=FALSE, upstream=1000, downstream=0)

## List extracted feature types
names(feat)
```

```

## Obtain feature lists by genes, here for promoter
split(feat$promoter, unlist(mcols(feat$promoter)$feature_by))

## Return all features in single GRanges object
unlist(feat)

## Not run:
## Sample from systemPipeRdata package
file <- system.file("extdata/annotation", "tair10.gff", package="systemPipeRdata")
txdb <- makeTxDbFromGFF(file=file, format="gff3", organism="Arabidopsis")
feat <- genFeatures(txdb, featuretype="all", reduce_ranges=FALSE, upstream=1000, downstream=0)

## End(Not run)

```

getQsubargs

Arguments for qsub

Description

Note: This function has been deprecated. Please use `clusterRun` instead. `getQsubargs` defines arguments to submit runX job(s) to queuing system (e.g. Torque) via `qsub`.

Usage

```
getQsubargs(software = "qsub", queue = "batch", Nnodes = "nodes=1", cores = as.numeric(gsub("^.* ",
```

Arguments

<code>software</code>	Software to use for submission to queuing system. Default is <code>qsub</code> .
<code>queue</code>	Name of queue to use. Default is <code>batch</code> .
<code>Nnodes</code>	Number of compute nodes to use for processing. Default is <code>nodes=1</code> .
<code>cores</code>	Number of CPU cores to use per compute node. Default will use what is provided by under <code>-p</code> in <code>myargs</code> of <code>systemArgs()</code> output.
<code>memory</code>	Amount of RAM to reserve per node.
<code>time</code>	Walltime limit each job is allowed to run per node.

Value

list

Author(s)

Thomas Girke

Examples

```

## Construct SYSargs object from param and targets files
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
args
names(args); modules(args); cores(args); outpaths(args); sysargs(args)

## Not run:
## Execute SYSargs on single machine
runCommandLine(args=args)

## Execute SYSargs on multiple machines
qsubargs <- getQsubargs(queue="batch", Nnodes="nodes=1", cores=cores(tophat), memory="mem=10gb", time="wallt")
qsubRun(args=args, qsubargs=qsubargs, Nqsubs=1, package="systemPipeR")
## Alignment stats
read_statsDF <- alignStats(fqpaths=tophatargs$infile1, bampaths=bampaths, fqgz=TRUE)
read_statsDF <- cbind(read_statsDF[targets$FileName,], targets)
write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE, sep="\t")

## End(Not run)

```

GOHyperGAll

GO term enrichment analysis for large numbers of gene sets

Description

To test a sample population of genes for over-representation of GO terms, the core function GOHyperGAll computes for all nodes in the three GO networks (BP, CC and MF) an enrichment test based on the hypergeometric distribution and returns the corresponding raw and Bonferroni corrected p-values. Subsequently, a filter function supports GO Slim analyses using default or custom GO Slim categories. Several convenience functions are provided to process large numbers of gene sets (e.g. clusters from partitioning results) and to visualize the results.

Note: GOHyperGAll provides similar utilities as the GOHyperG function in the GOstats package. The main difference is that GOHyperGAll simplifies processing of large numbers of gene sets, as well as the usage of custom array-to-gene and gene-to-GO mappings.

Usage

```

## Generate gene-to-GO mappings and store as catDB object
makeCATdb(myfile, lib = NULL, org = "", colno = c(1, 2, 3), idconv = NULL, rootUK=FALSE)

## Enrichment function
GOHyperGAll(catdb, gocat = "MF", sample, Nannot = 2)

## GO slim analysis
GOHyperGAll_Subset(catdb, GOHyperGAll_result, sample = test_sample, type = "goSlim", myslimv)

## Reduce GO term redundancy
GOHyperGAll_Simplify(GOHyperGAll_result, gocat = "MF", cutoff = 0.001, correct = TRUE)

## Batch analysis of many gene sets

```

```
GOCluster_Report(catdb, setlist, id_type = "affy", method = "all", CLSZ = 10, cutoff = 0.001, gocats
```

```
## Bar plot of GOCluster_Report results
goBarplot(GOBatchResult, gocat)
```

Arguments

myfile	File with gene-to-GO mappings. Sample files can be downloaded from geneontology.org (http://geneontology.org/GO.downloads.annotations.shtml) or from BioMart as shown in example below.
colno	Column numbers referencing in myfile the three target columns containing GOID, GeneID and GOCAT, in that order.
org	Optional argument. Currently, the only valid option is org="Arabidopsis" to get rid of transcript duplications in this particular annotation.
lib	If the gene-to-GO mappings are obtained from a *.db package from Bioconductor then the package name can be specified under the lib argument of the sampleDFgene2GO function.
idconv	Optional id conversion data.frame
catdb	catdb object storing mappings of genes to annotation categories. For details, see ?"SYSargs-class".
rootUK	If the argument rootUK is set to TRUE then the root nodes are treated as terminal nodes to account for the new unknown terms.
sample	character vector containing the test set of gene identifiers
Nannot	Defines the minimum number of direct annotations per GO node from the sample set to determine the number of tested hypotheses for the p-value adjustment.
gocat	Specifies the GO type, can be assigned one of the following character values: "MF", "BP" and "CC".
GOHyperGAll_result	data.frame generated by GOHyperGAll
type	The function GOHyperGAll_Subset subsets the GOHyperGAll results by directly assigned GO nodes or custom goSlim categories. The argument type can be assigned the values goSlim or assigned.
myslimv	optional argument to provide custom goSlim vector
cutoff	p-value cutoff for GO terms to show in result data.frame
correct	If TRUE the function will favor the selection of terminal (informationich) GO terms that have at the same time a large number of sample matches.
setlist	list of character vectors containing gene IDs (or array feature IDs). The names of the list components correspond to the set labels, e.g. DEG comparisons or cluster IDs.
id_type	specifies type of IDs in input, can be assigned gene or affy
method	Specifies analysis type. Current options are all for GOHyperGAll, slim for GOHyperGAll_Subset or simplify for GOHyperGAll_Simplify.
CLSZ	minimum gene set (cluster) size to consider. Gene sets below this cutoff will be ignored.
gocats	Specifies GO type, can be assigned the values "MF", "BP" and "CC".
recordSpecGO	argument to report in the result data.frame specific GO IDs for any of the 3 ontologies disregarding whether they meet the specified p-value cutoff, e.g: recordSpecGO=c("GO:0003674", "GO:0008150", "GO:0005575")

```
GOBatchResult  data.frame generated by GOCluster_Report
...           additional arguments to pass on
```

Details

GOHyperGAll_Simplify: The result data frame from GOHyperGAll will often contain several connected GO terms with significant scores which can complicate the interpretation of large sample sets. To reduce this redundancy, the function GOHyperGAll_Simplify subsets the data frame by a user specified p-value cutoff and removes from it all GO nodes with overlapping children sets (OFFSPRING), while the best scoring nodes are retained in the result data.frame.

GOCluster_Report: performs the three types of GO term enrichment analyses in batch mode: GOHyperGAll, GOHyperGAll_Subset or GOHyperGAll_Simplify. It processes many gene sets (e.g. gene expression clusters) and returns the results conveniently organized in a single result data frame.

Value

makeCATdb generates catDB object from file.

Author(s)

Thomas Girke

References

This workflow has been published in Plant Physiol (2008) 147, 41-57.

See Also

GOHyperGAll_Subset, GOHyperGAll_Simplify, GOCluster_Report, goBarplot

Examples

```
## Not run:

## Obtain annotations from BioMart
listMarts() # To choose BioMart database
m <- useMart("ENSEMBL_MART_PLANT"); listDatasets(m)
m <- useMart("ENSEMBL_MART_PLANT", dataset="athaliana_eg_gene")
listAttributes(m) # Choose data types you want to download
go <- getBM(attributes=c("go_accession", "tair_locus", "go_namespace_1003"), mart=m)
go <- go[go[,3]!="",]; go[,3] <- as.character(go[,3])
write.table(go, "GOannotationsBiomart_mod.txt", quote=FALSE, row.names=FALSE, col.names=FALSE, sep="\t")

## Create catDB instance (takes a while but needs to be done only once)
catdb <- makeCATdb(myfile="GOannotationsBiomart_mod.txt", lib=NULL, org="", colno=c(1,2,3), idconv=NULL)
catdb

## Create catDB from Bioconductor annotation package
# catdb <- makeCATdb(myfile=NULL, lib="ath1121501.db", org="", colno=c(1,2,3), idconv=NULL)

## AffyID-to-GeneID mappings when working with AffyIDs
# affy2locusDF <- systemPipeR::.AffyID2GeneID(map = "ftp://ftp.arabidopsis.org/home/tair/Microarrays/Affymetrix")
# catdb_conv <- makeCATdb(myfile="GOannotationsBiomart_mod.txt", lib=NULL, org="", colno=c(1,2,3), idconv=list())
# systemPipeR::.AffyID2GeneID(catdb=catdb_conv, affyIDs=c("244901_at", "244902_at"))
```



```

## Next time catDB can be loaded from file
save(catdb, file="catdb.RData")
load("catdb.RData")

## Perform enrichment test on single gene set
test_sample <- unique(as.character(catmap(catdb)$D_MF[1:100,"GeneID"]))
GOHyperGAll(catdb=catdb, gocat="MF", sample=test_sample, Nannot=2)[1:20,]

## GO Slim analysis by subsetting results accordingly
GOHyperGAll_result <- GOHyperGAll(catdb=catdb, gocat="MF", sample=test_sample, Nannot=2)
GOHyperGAll_Subset(catdb, GOHyperGAll_result, sample=test_sample, type="goSlim")

## Reduce GO term redundancy in 'GOHyperGAll_results'
simplifyDF <- GOHyperGAll_Simplify(GOHyperGAll_result, gocat="MF", cutoff=0.001, correct=T)
# Returns the redundancy reduced data set.
data.frame(GOHyperGAll_result[GOHyperGAll_result[,1]]

## Batch Analysis of Gene Clusters
testlist <- list(Set1=test_sample)
GOBatchResult <- GOCluster_Report(catdb=catdb, setlist=testlist, method="all", id_type="gene", CLSZ=10, cutof

## Plot 'GOBatchResult' as bar plot
goBarplot(GOBatchResult, gocat="MF")

## End(Not run)

```

INTERSECTset-class *Class "INTERSECTset"*

Description

Container for storing standard intersect results created by the `overLapper` function. The `setlist` slot stores the original label sets as vectors in a list; `intersectmatrix` organizes the label sets in a present-absent matrix; `complexitylevels` represents the number of comparisons considered for each comparison set as vector of integers; and `intersectlist` contains the standard intersect vectors.

Objects from the Class

Objects can be created by calls of the form `new("INTERSECTset", ...)`.

Slots

`setlist`: Object of class "list": list of vectors
`intersectmatrix`: Object of class "matrix": binary matrix
`complexitylevels`: Object of class "integer": vector of integers
`intersectlist`: Object of class "list": list of vectors

Methods

as.list signature(x = "INTERSECTset"): coerces INTERSECTset to list
coerce signature(from = "list", to = "INTERSECTset"): as(list, "INTERSECTset")
complexitylevels signature(x = "INTERSECTset"): extracts data from complexitylevels slot
intersectlist signature(x = "INTERSECTset"): extracts data from intersectlist slot
intersectmatrix signature(x = "INTERSECTset"): extracts data from intersectmatrix slot
length signature(x = "INTERSECTset"): returns number of original label sets
names signature(x = "INTERSECTset"): extracts slot names
setlist signature(x = "INTERSECTset"): extracts data from setlist slot
show signature(object = "INTERSECTset"): summary view of INTERSECTset objects

Author(s)

Thomas Girke

See Also

overLapper, vennPlot, olBarplot, VENNset-class

Examples

```
showClass("INTERSECTset")

## Sample data
setlist <- list(A=sample(letters, 18), B=sample(letters, 16),
               C=sample(letters, 20), D=sample(letters, 22),
               E=sample(letters, 18), F=sample(letters, 22))

## Create VENNset
interset <- overLapper(setlist[1:5], type="intersects")
class(interset)

## Accessor methods for VENNset/INTERSECTset objects
names(interset)
setlist(interset)
intersectmatrix(interset)
complexitylevels(interset)
intersectlist(interset)

## Coerce VENNset/INTERSECTset object to list
as.list(interset)
```

Description

The constructor functions create an SYSargs2 S4 class object from three input files: a CWL param and input files, and one simple tabular or yml file, a targets file. The latter is optional for workflow steps lacking input files. The CWL param provides all the parameters required for running command-line software, following the standard and specification defined on [Common Workflow Language \(CWL\)](#). The input file provides additional information for the command-line, allowing each sample level input/outfile operation uses its own SYSargs2 instance. In the targets file users could provide the paths to the initial sample input files (e.g. FASTQ) along with sample labels, and if appropriate biological replicate and contrast information for controlling differential abundance analyses.

Usage

```
loadWorkflow(targets = NULL, wf_file, input_file, dir_path = ".")
```

Arguments

targets	either the path to targets file or an object of SYSargs2 class. The targets file can be either a simple tabular or yml file. Also, it is possible to assign NULL to run the pipeline without the 'targets' file. This can be useful for running specific workflows that do not require input files.
wf_file	name and path to CWL param file.
input_file	name and path to input file.
dir_path	full path to the directory with the CWL param and input files.

Value

SYSargs2 object

Author(s)

Daniela Cassol and Thomas Girke

See Also

```
renderWF showClass("SYSargs2")
```

Examples

```
## Construct SYSargs2 object from CWL param, CWL input, and targets files
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF
```

mergeBamByFactor	<i>Merge BAM files based on factor</i>
------------------	--

Description

Merges BAM files based on sample groupings provided by a factor using internally the mergeBam function from the Rsamtools package. The function also returns an updated SYSargs or SYSargs2 object containing the paths to the merged BAM files as well as to the unmerged BAM files if there are any. All rows of merged parent samples are removed.

The functionality provided by mergeBamByFactor is useful for experiments where pooling of replicates is advantageous to maximize the depth of read coverage, such as prior to peak calling in ChIP-Seq or miRNA gene prediction experiments.

Usage

```
mergeBamByFactor(args, mergefactor = "Factor", overwrite = FALSE, silent = FALSE, ...)
```

Arguments

args	An instance of SYSargs or SYSargs2 constructed from a targets file where the first column (targetsin(args) or targets.as.df(targets(args))) contains the paths to the BAM files along with the column title FileName.
mergefactor	factor containing the grouping information required for merging the BAM files referenced in the first column of targetsin(args) or targets.as.df(targets(args)). The default uses Factor column from the targets files as factor. The latter merges BAM files for which replicates are specified in the Factor column.
overwrite	If overwrite=FALSE existing BAM files of the same name will not be overwritten.
silent	If silent=TRUE print statements will be suppressed.
...	To pass on additional arguments to the internally used mergeBam function from Rsamtools.

Value

The merged BAM files will be written to output files with the following naming convention: <first_BAM_file_name>_<grouping>. In addition, the function returns an updated SYSargs or SYSargs2 object where all output file paths contain the paths to the merged BAM files. The rows of the merged parent samples are removed and the rows of the unmerged samples remain unchanged.

Author(s)

Thomas Girke

See Also

writeTargetsout, writeTargetsRef

Examples

```
## Construct initial SYSargs object
targetspath <- system.file("extdata", "targets_chip.txt", package="systemPipeR")
parampath <- system.file("extdata", "bowtieSE.param", package="systemPipeR")
args <- systemArgs(sysma=parampath, mytargets=targetspath)

## Not run:
## After running alignmets (e.g. with Bowtie2) generate targets file
## for the corresponding BAM files. The alignment step is skipped here.
writeTargetsout(x=args, file="targets_bam.txt", overwrite=TRUE)
args <- systemArgs(sysma=NULL, mytargets="targets_bam.txt")

## Merge BAM files and return updated SYSargs object
args_merge <- mergeBamByFactor(args, overwrite=TRUE, silent=FALSE)

## Export modified targets file
writeTargetsout(x=args_merge, file="targets_mergeBamByFactor.txt", overwrite=TRUE)

## End(Not run)
```

 module

Interface to allow full use of the Environment Modules system for Unix

Description

The function `module` enables use of the Environment Modules system (<http://modules.sourceforge.net/>) from within the R environment. By default the user's login shell environment (ie. `bash -l`) will be used to initialize the current session. The module function can also; load or unload specific software, list all the loaded software within the current session, and list all the applications available for loading from the module system. Lastly, the module function can remove all loaded software from the current session.

Usage

```
module(action_type, module_name="")
```

Arguments

<code>action_type</code>	Name of the action to be executed as character vector. The following switches are accepted: <code>avail</code> , <code>list</code> , <code>init</code> , <code>load</code> , <code>unload</code> , and <code>clear</code> .
<code>module_name</code>	Name of software to load as character vector.

Author(s)

Jordan Hayes and Thomas Girke

Examples

```
## Not run:
## List all available software from the module system
module("avail")

## List loaded software in the current session
```

```
module("list")

## Example for loading a software into the shell environment
module("load","tophat")

## Example for removing software from the shell environment
module("unload", "tophat")

## Clear all of the software from the shell's initialization files
module("clear")

## List and load all the software loaded in users default login shell into the current session (default)
module("init")

## End(Not run)
```

moduleload

Interface to module system

Description

Functions to list and load software from a module system in R. The functions are the equivalent of module avail and module load on the Linux command-line, respectively.

Usage

```
moduleload(module, envir="PATH")

modulelist()
```

Arguments

module	Name of software to load character vector.
envir	One or many environment variables passed on as character vector.

Author(s)

Tyler Backman and Thomas Girke

Examples

```
## Not run:
## List all software from module system
modulelist()

## Examples for loading software from module system
moduleload(module="bowtie2/2.0.6", envir="PATH")
moduleload(module="python", envir=c("PATH", "LD_LIBRARY_PATH", "PYTHONPATH"))

## End(Not run)
```

olBarplot

*Bar plot for intersect sets***Description**

Generates bar plots of the intersect counts of VENNset and INTERSECTset objects generated by the overLapper function. It is an alternative to Venn diagrams (e.g. vennPlot) that scales to larger numbers of label sets. By default the bars in the plot are colored and grouped by complexity levels of the intersect sets.

Usage

```
olBarplot(x, mincount = 0, complexity="default", myxlabel = "default", myylabel="Counts", mytitle =
```

Arguments

x	Object of class VENNset or INTERSECTset.
mincount	Sets minimum number of counts to consider in the bar plot. Default mincount=0 considers all counts.
complexity	Allows user to limit the bar plot to specific complexity levels of intersects by specifying the chosen ones with an integer vector. Default complexity="default" considers all complexity levels.
myxlabel	Defines label of x-axis.
myylabel	Defines label of y-axis.
mytitle	Defines main title of plot.
...	Allows to pass on additional arguments to geom_bar from ggplot2. For instance, fill=seq(along=vennlist(x)) or fill=seq(along=intersectlist(x)) will assign a different color to each bar, or fill="blue" will color all of them blue. The default bar coloring is by complexity levels of the intersect sets.

Value

Bar plot.

Note

The functions provided here are an extension of the Venn diagram resources on this site: <http://manuals.bioinformatics.ucr.edu/home/venn/Venn-Diagrams>

Author(s)

Thomas Girke

See Also

overLapper, vennPlot

Examples

```

## Sample data: list of vectors with object labels
setlist <- list(A=sample(letters, 18), B=sample(letters, 16),
               C=sample(letters, 20), D=sample(letters, 22),
               E=sample(letters, 18), F=sample(letters, 22))

## 2-way Venn diagram
vennset <- overLapper(setlist[1:2], type="vennsets")
vennPlot(vennset)

## 3-way Venn diagram
vennset <- overLapper(setlist[1:3], type="vennsets")
vennPlot(vennset)

## 4-way Venn diagram
vennset <- overLapper(setlist[1:4], type="vennsets")
vennPlot(list(vennset, vennset))

## Pseudo 4-way Venn diagram with circles
vennPlot(vennset, type="circle")

## 5-way Venn diagram
vennset <- overLapper(setlist[1:5], type="vennsets")
vennPlot(vennset)

## Alternative Venn count input to vennPlot (not recommended!)
counts <- sapply(vennlist(vennset), length)
vennPlot(counts)

## 6-way Venn comparison as bar plot
vennset <- overLapper(setlist[1:6], type="vennsets")
olBarplot(vennset, mincount=1)

## Bar plot of standard intersect counts
intersect <- overLapper(setlist, type="intersects")
olBarplot(intersect, mincount=1)

## Accessor methods for VENNset/INTERSECTset objects
names(vennset)
names(intersect)
setlist(vennset)
intersectmatrix(vennset)
complexitylevels(vennset)
vennlist(vennset)
intersectlist(intersect)

## Coerce VENNset/INTERSECTset object to list
as.list(vennset)
as.list(intersect)

## Pairwise intersect matrix and heatmap
olMA <- sapply(names(setlist),
              function(x) sapply(names(setlist),
                                function(y) sum(setlist[[x]] %in% setlist[[y]])))
olMA
heatmap(olMA, Rowv=NA, Colv=NA)

```



```
## Presence-absence matrices for large numbers of sample sets
interaset <- overLapper(setlist=setlist, type="intersects", complexity=2)
(paMA <- intersectmatrix(interaset))
heatmap(paMA, Rowv=NA, Colv=NA, col=c("white", "gray"))
```

olRanges

Identify Range Overlaps for IRanges and GRanges Object

Description

Function for identifying consensus peak among two peaks sets sharing a minimum relative overlap.

Usage

```
olRanges(query, subject, output = "gr")
```

Arguments

query	Object of class GRanges, which is a vector of genomic locations and associated annotations.
subject	Object of class GRanges.
output	By default "gr" returns any overlap with OL length information in an object of class GRanges. Also, can returns an object of class data.frame with "df".

Author(s)

Thomas Girke

Examples

```
## Sample Data Sets
grq <- GRanges(seqnames = Rle(c("chr1", "chr2", "chr1", "chr3"), c(1, 3, 2, 4)), ranges = IRanges(seq(1, 100, by = 10), width = 10))
grs <- shift(grq[c(2,5,6)], 5)
## Run olRanges function
olRanges(query=grq, subject=grs, output="df")
olRanges(query=grq, subject=grs, output="gr")
```

output_update

Updates the output files paths in the SYSargs2 object

Description

After executing all the command-lines by the runCommandline function, the output files can be created in specific directories rather than results in a particular directory. Also, the runCommandline function allows converting the SAM file outputs to sorted and indexed BAM files. Thus, the output_update function allows updating the location of these files in the output of the SYSargs2 object.

Usage

```
output_update(args, dir = TRUE, dir.name = NULL, replace = FALSE, extension = NULL)
```

Arguments

args	object of class SYSargs2.
dir	assign TRUE to update the location of the output files in the args object accordingly with the workflow name directory. Default is dir=TRUE.
dir.name	if the results directory name is not specified in the input file, it is possible to specify here the name. This argument is required if the path name return NULL from the input file. Default is dir.name=NULL.
replace	replace the extension for selected output files in the args object. Default is replace=FALSE.
extension	object of class "character" storing the current extension of the files and the respective replacement. For example, runCommandline function by default autotetects SAM file outputs in the args object and create the BAM files. In order to update the output of args object, the extension argument should be set: extension = c(".sam", ".bam").

Value

SYSargs2 object with output location files updated.

Author(s)

Daniela Cassol and Thomas Girke

See Also

To check directory name in the input file: `yamlinput(WF)$results_path$path`.

Examples

```
## Construct SYSargs2 object from CWL param, CWL input, and targets files
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF
output(WF)

## Not run:
runCommandline(args=WF, make_bam=TRUE)
## Output paths update
WF <- output_update(WF, dir=FALSE, replace=TRUE, extension=c(".sam", ".bam"))

runCommandline(args=WF, make_bam=TRUE, dir=TRUE)
## Output paths update
WF <- output_update(WF, dir=TRUE, replace=TRUE, extension=c(".sam", ".bam"))

## End(Not run)
```

Description

Function for computing Venn intersects or standard intersects among large numbers of label sets provided as list of vectors. The resulting intersect objects can be used for plotting 2-5 way Venn diagrams or intersect bar plots using the functions `vennPlot` or `olBarplot`, respectively. The `overLapper` function scales to 2-20 or more label vectors for Venn intersect calculations and to much larger sample numbers for standard intersects. The different intersect types are explained below under the definition of the `type` argument. The upper Venn limit around 20 label sets is unavoidable because the complexity of Venn intersects increases exponentially with the label set number n according to this relationship: $2^n - 1$. The current implementation of the plotting function `vennPlot` supports Venn diagrams for 2-5 label sets. To visually analyze larger numbers of label sets, a variety of intersect methods are introduced in the `olBarplot` help file. These methods are much more scalable than Venn diagrams, but lack their restrictive intersect logic.

Usage

```
overLapper(setlist, complexity = "default", sep = "_", cleanup = FALSE, keepdups = FALSE, type)
```

Arguments

<code>setlist</code>	Object of class <code>list</code> where each list component stores a label set as vector and the name of each label set is stored in the name slot of each list component. The names are used for naming the label sets in all downstream analysis steps and plots.
<code>complexity</code>	Complexity level of intersects specified as integer vector. For Venn intersects it needs to be assigned <code>1:length(setlist)</code> (default). If <code>complexity=2</code> the function returns all pairwise intersects.
<code>sep</code>	Character used to separate set labels.
<code>cleanup</code>	If set to <code>TRUE</code> then all characters of the label sets are set to upper case, and leading and trailing spaces are removed. The default <code>cleanup=FALSE</code> omits this step.
<code>keepdups</code>	By default all duplicates are removed from the label sets. The setting <code>keepdups=TRUE</code> will retain duplicates by appending a counter to each entry.
<code>type</code>	With the default setting <code>type="vennsets"</code> the <code>overLapper</code> function computes the typical Venn intersects for the label sets provided under <code>setlist</code> . With the setting <code>type="intersects"</code> the function will compute pairwise intersects (not compatible with Venn diagrams). Venn intersects follow the typical 'only in' intersect logic of Venn comparisons, such as: labels present only in set A, labels present only in the intersect of A & B, etc. Due to this restrictive intersect logic, the combined Venn sets contain no duplicates. In contrast to this, regular intersects follow this logic: labels present in the intersect of A & B, labels present in the intersect of A & B & C, etc. This approach results usually in many duplications of labels among the intersect sets.

Details

Additional Venn diagram resources are provided by the packages `limma`, `gplots`, `vennerable`, `eVenn` and `VennDiagram`, or online resources such as `shapes`, `Venn Diagram Generator` and `Venny`.

Value

overLapper returns standard intersect and Venn intersect results as INTERSECTset or VENNset objects, respectively. These S4 objects contain the following components:

```
setlist          Original label sets accessible with setlist().
intersectmatrix Present-absent matrix accessible with intersectmatrix(), where each overlap
                 set in the vennlist data component is labeled according to the label set names
                 provided under setlist. For instance, the composite name 'ABC' indicates
                 that the entries are restricted to A, B and C. The separator used for naming the
                 intersect sets can be specified under the sep argument.
complexitylevels Complexity levels accessible with complexitylevels().
vennlist         Venn intersects for VENNset objects accessible with vennlist().
intersectlist    Standard intersects for INTERSECTset objects accessible with intersectlist().
```

Note

The functions provided here are an extension of the Venn diagram resources on this site: <http://manuals.bioinformatics.ucr.edu/home/1/Venn-Diagrams>

Author(s)

Thomas Girke

References

See examples in 'The Electronic Journal of Combinatorics': <http://www.combinatorics.org/files/Surveys/ds5/VennSymm>

See Also

vennPlot, olBarplot

Examples

```
## Sample data
setlist <- list(A=sample(letters, 18), B=sample(letters, 16),
               C=sample(letters, 20), D=sample(letters, 22),
               E=sample(letters, 18), F=sample(letters, 22))

## 2-way Venn diagram
vennset <- overLapper(setlist[1:2], type="vennsets")
vennPlot(vennset)

## 3-way Venn diagram
vennset <- overLapper(setlist[1:3], type="vennsets")
vennPlot(vennset)

## 4-way Venn diagram
vennset <- overLapper(setlist[1:4], type="vennsets")
vennPlot(list(vennset, vennset))

## Pseudo 4-way Venn diagram with circles
vennPlot(vennset, type="circle")
```

```

## 5-way Venn diagram
vennset <- overLapper(setlist[1:5], type="vennsets")
vennPlot(vennset)

## Alternative Venn count input to vennPlot (not recommended!)
counts <- sapply(vennlist(vennset), length)
vennPlot(counts)

## 6-way Venn comparison as bar plot
vennset <- overLapper(setlist[1:6], type="vennsets")
olBarplot(vennset, mincount=1)

## Bar plot of standard intersect counts
intersect <- overLapper(setlist, type="intersects")
olBarplot(intersect, mincount=1)

## Accessor methods for VENNset/INTERSECTset objects
names(vennset)
names(intersect)
setlist(vennset)
intersectmatrix(vennset)
complexitylevels(vennset)
vennlist(vennset)
intersectlist(intersect)

## Coerce VENNset/INTERSECTset object to list
as.list(vennset)
as.list(intersect)

## Pairwise intersect matrix and heatmap
olMA <- sapply(names(setlist),
function(x) sapply(names(setlist),
function(y) sum(setlist[[x]] %in% setlist[[y]])))
olMA
heatmap(olMA, Rowv=NA, Colv=NA)

## Presence-absence matrices for large numbers of sample sets
intersect <- overLapper(setlist=setlist, type="intersects", complexity=2)
(paMA <- intersectmatrix(intersect))
heatmap(paMA, Rowv=NA, Colv=NA, col=c("white", "gray"))

```

plotfeatureCoverage *Plot feature coverage results*

Description

Plots the 3 tabular data types (A-C) generated by the featureCoverage function. It accepts data from single or many features (e.g. CDSs) and samples (BAM files). The coverage from multiple features will be summarized using methods such as mean, while the data from multiple samples will be plotted in separate panels.

Usage

```
plotfeatureCoverage(covMA, method = mean, scales = "fixed", extendylim=2, scale_count_val = 10^6)
```

Arguments

covMA	Object of class <code>data.frame</code> generated by <code>featureCoverage</code> function.
method	Defines the summary statistics to use when <code>covMA</code> contains coverage data from multiple features (e.g. transcripts). The default calculates the mean coverage for each position and/or bin of the corresponding coverage vectors.
scales	Scales setting passed on to the <code>facet_wrap</code> function of <code>ggplot2</code> . For details see <code>ggplot2::facet_wrap</code> . The default <code>fixed</code> assures a constant scale across all bar plot panels, while <code>free</code> uses the optimum scale within each bar plot panel. To evaluate plots in all their details, it may be necessary to generate two graphics files one for each scaling option.
extendylim	Allows to extend the upper limit of the y axis when <code>scales=fixed</code> . Internally, the function identifies the maximum value in the data and then multiplies this maximum value by the value provided under <code>extendylim</code> . The default is set to <code>extendylim=2</code> .
scale_count_val	Scales (normalizes) the read counts to a fixed value of aligned reads in each sample such as counts per million aligned reads (default is 10^6). For this calculation the <code>N_total_aligned</code> values are used that are reported in the input <code>data.frame</code> generated by the upstream <code>featureCoverage</code> function. Assign <code>NULL</code> to turn off scaling.

Value

Currently, the function returns `ggplot2` bar plot graphics.

Author(s)

Thomas Girke

See Also

`featureCoverage`

Examples

```
## Construct SYSargs object from param and targets files
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)

## Not run:
## Features from sample data of systemPipeRdata package
library(GenomicFeatures)
file <- system.file("extdata/annotation", "tair10.gff", package="systemPipeRdata")
txdb <- makeTxDbFromGFF(file=file, format="gff3", organism="Arabidopsis")

## (A) Generate binned coverage for two BAM files and 4 transcripts
gr1 <- cdsBy(txdb, "tx", use.names=TRUE)
fcov <- featureCoverage(bfl=BamFileList(outpaths(args)[1:2]), gr1=gr1[1:4], resizereads=NULL,
  readlengthrange=NULL, Nbins=20, method=mean, fixedmatrix=FALSE,
  resizefeatures=TRUE, upstream=20, downstream=20)
fcov <- featureCoverage(bfl=BamFileList(outpaths(args)[1:2]), gr1=gr1[1:4], resizereads=NULL,
  readlengthrange=NULL, Nbins=20, method=mean, fixedmatrix=TRUE,
```

```

resizefeatures=TRUE, upstream=20, downstream=20)
plotfeatureCoverage(covMA=fcov, method=mean, scales="fixed", scale_count_val=10^6)

## (B) Coverage matrix upstream and downstream of start/stop codons
fcov <- featureCoverage(bfl=BamFileList(outpaths(args)[1:2]), grl=grl[1:4], resizereads=NULL,
  readlengthrange=NULL, Nbins=NULL, method=mean, fixedmatrix=TRUE,
  resizefeatures=TRUE, upstream=20, downstream=20)
plotfeatureCoverage(covMA=fcov, method=mean, scales="fixed", scale_count_val=10^6)

## (C) Combined matrix for both binned and start/stop codon
fcov <- featureCoverage(bfl=BamFileList(outpaths(args)[1:2]), grl=grl[1:4], resizereads=NULL,
  readlengthrange=NULL, Nbins=20, method=mean, fixedmatrix=TRUE,
  resizefeatures=TRUE, upstream=20, downstream=20, outfile="results/test.xls")
plotfeatureCoverage(covMA=fcov, method=mean, scales="fixed", scale_count_val=10^6)

## (D) Rle coverage objects one for each query feature
fcov <- featureCoverage(bfl=BamFileList(outpaths(args)[1:2]), grl=grl[1:4], resizereads=NULL,
  readlengthrange=NULL, Nbins=NULL, method=mean, fixedmatrix=FALSE,
  resizefeatures=TRUE, upstream=20, downstream=20)

## End(Not run)

```

plotfeaturetypeCounts *Plot read distribution across genomic features*

Description

Function to visualize the distribution of reads across different feature types for many alignment files in parallel. The plots are stacked bar plots representing the raw or normalized read counts for the sense and antisense strand of each feature. The graphics results are generated with ggplot2. Typically, the expected input is generated with the affiliated featuretypeCounts function.

Usage

```
plotfeaturetypeCounts(x, graphicsfile, graphicsformat = "pdf", scales = "fixed", anyreadlength = FALSE,
  drop_N_total_aligned = TRUE, scale_count_val = 10^6, scale_length_val = NULL)
```

Arguments

x	data.frame with feature counts generated by the featuretypeCounts function.
graphicsfile	Path to file where to write the output graphics. Note, the function returns the graphics instructions from ggplot2 for interactive plotting in R. However, due to the complexity of the graphics generated here, the finished results are written to a file directly.
graphicsformat	Graphics file format. Currently, supported formats are: pdf, png or jpeg. Argument accepts one of them as character string.
scales	Scales setting passed on to the facet_wrap function of ggplot2. For details see ggplot2::facet_wrap. The default fixed assures a constant scale across all bar plot panels, while free uses the optimum scale within each bar plot panel. To evaluate plots in all their details, it may be necessary to generate two graphics files one for each scaling option.

- anyreadlength** If set to TRUE read length specific read counts will be summed up to a single count value to plot read counts for any read length. Otherwise the bar plots will show the counts for each read length value.
- drop_N_total_aligned**
If set to TRUE the special feature count `N_total_aligned` will not be included as a separate feature in the plots. However, the information will still be used internally for scaling the read counts to a fixed value if this option is requested under the `scale_count_val` argument.
- scale_count_val**
Scales (normalizes) the read counts to a fixed value of aligned reads in each sample such as counts per million aligned reads (default is 10^6). For this calculation the `N_total_aligned` values are used that are reported in the input `data.frame` generated by the upstream `featuretypeCounts` function. Assign NULL to turn off scaling by aligned reads.
- scale_length_val**
Allows to adjust the raw or scaled read counts to a constant length interval (e.g. `scale_length_val=10^3` in bps) considering the total genomic length of the corresponding feature type. The required genomic length information for each feature type is obtained from the `FeaturetypeLength` column of the input `data.frame` generated by the `featuretypeCount` function. To turn off feature length adjustment, assign NULL (default).

Value

The function returns bar plot graphics for aligned read counts with read length resolution if the input contains this information and argument `anyreadlength` is set to FALSE. If the input contains counts for any read length and/or `anyreadlength=TRUE` then there will be only one bar per feature and sample. Due to the complexity of the plots, the results are directly written to file in the chosen graphics format. However, the function also returns the plotting instructions returned by `ggplot2` to display the result components using R's plotting device.

Author(s)

Thomas Girke

See Also

`featuretypeCounts`, `genFeatures`

Examples

```
## Construct SYSargs object from param and targets files
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)

## Not run:
## Features from sample data of systemPipeRdata package
library(GenomicFeatures)
file <- system.file("extdata/annotation", "tair10.gff", package="systemPipeRdata")
txdb <- makeTxDbFromGFF(file=file, format="gff3", organism="Arabidopsis")
feat <- genFeatures(txdb, featuretype="all", reduce_ranges=TRUE, upstream=1000, downstream=0, verbose=TRUE)

## Generate and plot feature counts for specific read lengths
```



```

fc <- featuretypeCounts(bfl=BamFileList(outpaths(args), yieldSize=50000), grl=feat, singleEnd=TRUE, readlength=feat)
p <- plotfeaturetypeCounts(x=fc, graphicsfile="featureCounts.pdf", graphicsformat="pdf", scales="fixed", any)

## Generate and plot feature counts for any read length
fc2 <- featuretypeCounts(bfl=BamFileList(outpaths(args), yieldSize=50000), grl=feat, singleEnd=TRUE, readlength=feat)
p2 <- plotfeaturetypeCounts(x=featureCounts2, graphicsfile="featureCounts2.pdf", graphicsformat="pdf", scales="fixed", any)

## End(Not run)

```

predORF

Predict ORFs

Description

Predicts open reading frames (ORFs) and coding sequences (CDSs) in DNA sequences provided as DNASTring or DNASTringSet objects.

Usage

```
predORF(x, n = 1, type = "gr1", mode = "orf", strand = "sense", longest_disjoint=FALSE, startcodon = "ATG", stopcodon = "TAA,TAG,TGA")
```

Arguments

x	DNA query sequence(s) provided as DNASTring or DNASTringSet object.
n	Defines the maximum number of ORFs to return for each input sequence. The ORFs identified are sorted decreasingly by their length. For instance, n=1 (default) returns the longest ORF, n=2 the two longest ones, and so on.
type	One of three options provided as character values: 'df' returns results as data.frame, while 'gr' and 'gr1' (default) return them as GRanges or GRangesList objects, respectively.
mode	The setting mode='ORF' returns a continuous reading frame that begins with a start codon and ends with a stop codon. The setting mode='CDS' return continuous reading frames that do not need to begin or end with start or stop codons, respectively.
strand	One of three options passed on as character vector of length one: 'sense' performs the predictions only for the sense strand of the query sequence(s), 'antisense' does it only for the antisense strand and 'both' does it for both strands.
longest_disjoint	If set to TRUE and n='all', the results will be subsetted to non-overlapping ORF set containing longest ORF.
startcodon	Defines the start codon(s) for ORF predictions. The default is set to the standard start codon 'ATG'. Any custom set of triplet DNA sequences can be assigned here.
stopcodon	Defines the stop codon(s) for ORF predictions. The default is set to the three standard stop codons 'TAA', 'TAG' and 'TGA'. Any custom set of triplet DNA sequences can be assigned here.

Value

Returns ORF/CDS ranges identified in query sequences as GRanges or data.frame object. The type argument defines which one of them will be returned. The objects contain the following columns:

- seqnames: names of query sequences
- subject_id: identified ORF/CDS ranges numbered by query
- start/end: start and end positions of ORF/CDS ranges
- strand: strand of query sequence used for prediction
- width: length of subject range in bases
- inframe2end: frame of identified ORF/CDS relative to 3' end of query sequence. This can be important if the query sequence was extracted directly upstream of an ORF (e.g. 5' UTR upstream of main ORF). The value 1 stands for in-frame with downstream ORF, while 2 or 3 indicates a shift of one or two bases, respectively.

Author(s)

Thomas Girke

See Also

scaleRanges

Examples

```
## Load DNA sample data set from Biostrings package
file <- system.file("extdata", "someORF.fa", package="Biostrings")
dna <- readDNASTringSet(file)

## Predict longest ORF for sense strand in each query sequence
(orf <- predORF(dna[1:4], n=1, type="gr", mode="orf", strand="sense"))

## Not run:
## Usage for more complex example
library(GenomicFeatures); library(systemPipeRdata)
gff <- system.file("extdata/annotation", "tair10.gff", package="systemPipeRdata")
txdb <- makeTxDbFromGFF(file=gff, format="gff3", organism="Arabidopsis")
futr <- fiveUTRsByTranscript(txdb, use.names=TRUE)
genome <- system.file("extdata/annotation", "tair10.fasta", package="systemPipeRdata")
dna <- extractTranscriptSeqs(FaFile(genome), futr)
uorf <- predORF(dna, n="all", mode="orf", longest_disjoint=TRUE, strand="sense")
grl_scaled <- scaleRanges(subject=futr, query=uorf, type="uORF", verbose=TRUE)
export.gff3(unlist(grl_scaled), "uorf.gff")

## End(Not run)
```

```
preprocessReads      Run custom read preprocessing functions
```

Description

Applies custom read preprocessing functions to single-end or paired-end FASTQ files. The function uses the `FastqStreamer` function from the `ShortRead` package to stream through large files in a memory-efficient manner.

Usage

```
preprocessReads(args, Fct, batchsize = 1e+05, overwrite = TRUE, ...)
```

Arguments

<code>args</code>	Object of class <code>SYSargs</code> or <code>SYSargs2</code> .
<code>Fct</code>	character string of custom read preprocessing function call where both the input and output needs to be an object of class <code>ShortReadQ</code> . The name of the input <code>ShortReadQ</code> object needs to be <code>fq</code> .
<code>batchsize</code>	Number of reads to process in each iteration by the internally used <code>FastqStreamer</code> function.
<code>overwrite</code>	If <code>TRUE</code> existing file will be overwritten.
<code>...</code>	To pass on additional arguments to the internally used <code>writeFastq</code> function.

Value

Writes to files in FASTQ format. Their names are specified by `outpaths(args)`.

Author(s)

Thomas Girke

See Also

`FastqStreamer`

Examples

```
## Preprocessing of single-end reads
param <- system.file("extdata", "trim.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
## Not run:
preprocessReads(args=args, Fct="trimLRPatterns(Rpattern='GCCCCGGTAA', subject=fq)", batchsize=100000, overwr

## End(Not run)

## Preprocessing of paired-end reads
param <- system.file("extdata", "trimPE.param", package="systemPipeR")
targets <- system.file("extdata", "targetsPE.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
## Not run:
```

```
preprocessReads(args=args, Fct="trimLRPatterns(Rpattern='GCCCGGTAA', subject=fq)", batchsize=100000, overwr
## End(Not run)
```

qsubRun

Submit command-line tools to cluster

Description

Note: This function has been deprecated. Please use `clusterRun` instead. `qsubRun` submits command-line tools to queue (e.g. Torque) of compute cluster using run specifications defined by `runX` and `getQsubargs` functions.

Usage

```
qsubRun(appfct="runCommandline(args=args, runid='01')", args, qsubargs, Nqsubs = 1, package = "systemPipeR")
```

Arguments

<code>appfct</code>	Accepts <code>runX</code> functions, such as <code>appfct="runCommandline(args,runid)"</code>
<code>args</code>	Argument list returned by <code>systemArgs()</code> .
<code>qsubargs</code>	Argument list returned by <code>getQsubargs()</code> .
<code>Nqsubs</code>	Integer defining the number of <code>qsub</code> processes. Note: the function will not assign more <code>qsub</code> processes than there are FASTQ files. E.g. if there are 10 FASTQ files and <code>Nqsubs=20</code> then the function will generate only 10 <code>qsub</code> processes. To increase the number of CPU cores used by each process, one can increase the <code>p</code> value under <code>systemArgs()</code> .
<code>package</code>	Package to load. Name provided as character vector of length one. Default is <code>systemPipeR</code> .
<code>shebang</code>	defines shebang (first line) used in submission shell script; default is set to <code>#!/bin/bash</code> .

Value

Returns list where list components contain FASTQ file names and their names are the `qsub` process IDs assigned by the queuing system. In addition, three files will be generated for each `qsub` submission process: `submitargs0X` (R object containing `appargs`), `submitargs0X.R` (R script using `appargs`) and `submitargs0X.sh` (shell submission script). In addition, the chosen `runX` function will output a `submitargs0X_log` file for each `qsub` process containing the executable commands processed by each `qsub` instance.

Author(s)

Thomas Girke

Examples

```

## Construct SYSargs object from param and targets files
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
args
names(args); modules(args); cores(args); outpaths(args); sysargs(args)

## Not run:
## Execute SYSargs on single machine
runCommandline(args=args)

## Execute SYSargs on multiple machines
qsubargs <- getQsubargs(queue="batch", Nnodes="nodes=1", cores=cores(tophat), memory="mem=10gb", time="wallt")
qsubRun(args=args, qsubargs=qsubargs, Nqsubs=1, package="systemPipeR")
## Alignment stats
read_statsDF <- alignStats(fqpaths=tophatargs$infile1, bampaths=bampaths, fqgz=TRUE)
read_statsDF <- cbind(read_statsDF[targets$FileName,], targets)
write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE, sep="\t")

## End(Not run)

```

readComp

Import sample comparisons from targets file

Description

Parses sample comparisons specified in <CMP> line(s) of targets file or in targetsheader slot of SYSargs object. All possible comparisons can be specified with 'CMPset: ALL'.

Usage

```
readComp(file, format = "vector", delim = "-")
```

Arguments

file	Path to targets file. Alternatively, a SYSargs or SYSargs2 object can be assigned.
format	Object type to return: vector or matrix.
delim	Delimiter to use when sample comparisons are returned as vector.

Value

list where each component is named according to the name(s) used in the <CMP> line(s) of the targets file. The list will contain as many sample comparisons sets (list components) as there are sample comparisons lines in the corresponding targets file.

Author(s)

Thomas Girke

Examples

```
## Return comparisons from targets file
targetspath <- system.file("extdata", "targets.txt", package="systemPipeR")
read.delim(targetspath, comment.char = "#")
readComp(file=targetspath, format="vector", delim="-")

## Return comparisons from SYSargs object
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
readComp(args, format = "vector", delim = "-")
```

renderWF

*Populate all the command-line in an SYSargs2 object***Description**

The SYSargs2 S4 class object is constructed from the loadWorkflow, which stores all the information and instructions needed for processing a set of input files with a specific command-line or a series of command-line within a workflow. The renderWF function populates all the command-line for each sample in each step of the particular workflow. Each sample level input/outfile operation uses its own SYSargs2 instance. The output of SYSargs2 define all the expected output files for each step in the workflow, which usually it is the sample input for the next step in an SYSargs2 instance. Between different instances, this connectivity is established by writing the subsetting output with the writeTargetsout function to a new targets file that serves as input to the next loadWorkflow and renderWF call. By chaining several SYSargs2 steps together one can construct complex workflows involving many sample-level input/output file operations with any combination of command-line or R-based software.

Usage

```
renderWF(WF, inputvars = c(FileName = "_FASTQ_PATH_"))
```

Arguments

WF	Object of class SYSargs2.
inputvars	variables list defined in the input file that matches the column names defined in the targets file.

Value

SYSargs2 object

Author(s)

Daniela Cassol and Thomas Girke

See Also

showClass("SYSargs2") loadWorkflow writeTargetsout

Examples

```
## Construct SYSargs2 object from CWL param, CWL input, and targets files
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF
```

returnRPKM

RPKM Normalization

Description

Converts read counts to RPKM normalized values.

Usage

```
returnRPKM(counts, ranges)
```

Arguments

counts	Count data frame, e.g. from an RNA-Seq experiment.
ranges	GRangesList object, e.g. generated by <code>exonsBy(txdb, by="gene")</code> .

Value

data.frame

Author(s)

Thomas Girke

Examples

```
## Not run:
countDFrpkM <- apply(countDF, 2, function(x) returnRPKM(counts=x, gffsub=eByg))

## End(Not run)
```

runCommandline *Execute SYSargs and SYSargs2*

Description

Function to execute system parameters specified in SYSargs and SYSargs2 object.

Usage

```
runCommandline(args, runid = "01", make_bam = TRUE, del_sam=TRUE, dir = FALSE, dir.name = NULL, force = FALSE)
```

Arguments

args	object of class SYSargs or SYSargs2.
runid	Run identifier used for log file to track system call commands. Default is "01".
make_bam	Auto detects SAM file outputs and converts them to sorted and indexed BAM files. Default is make_bam=TRUE.
del_sam	This option allows deleting the SAM files created when the make_BAM converts the SAM files to sorted and indexed BAM files. Default is del_sam=TRUE.
dir	This option allows creating an exclusive results folder for each step in the workflow and a sub-folder for each sample defined in the targets file. All the outputs and log files for the particular step will be created in the respective folders. Default is dir=FALSE. Option available only for an object of class SYSargs2.
dir.name	Name of the workflow directory. Default is dir.name=FALSE. Note: This argument is required when the dir=TRUE.
force	Internally, the function checks if the expected output files exist, and it skips the command lines when the respective files exist. If the argument force is set to TRUE, the command line will be executed and the files overwrite. Default is force=FALSE.
...	Additional arguments to pass on to runCommandline().

Value

Output files, their paths can be obtained with `outpaths()` from SYSargs container or `output()` from SYSargs2. In addition, a character vector is returned containing the same paths.

Author(s)

Daniela Cassol and Thomas Girke

Examples

```
#####
## Examples with \code{SYSargs} object ##
#####
## Construct SYSargs object from param and targets files
param <- system.file("extdata", "hisat2.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
args
```



```

names(args); modules(args); cores(args); outpaths(args); sysargs(args)

## Not run:
## Execute SYSargs on single machine
runCommandline(args=args)

## Execute SYSargs on multiple machines of a compute cluster.
file.copy(system.file("extdata", ".batchtools.conf.R", package="systemPipeR"), ".")
file.copy(system.file("extdata", "batchtools.slurm.tpl", package="systemPipeR"), ".")
resources <- list(walltime=120, ntasks=1, ncpus=cores(args), memory=1024)
reg <- clusterRun(args, FUN = runCommandline, conffile=".batchtools.conf.R", template="batchtools.slurm.tpl")

## Monitor progress of submitted jobs
getStatus(reg=reg)
file.exists(outpaths(args))

## Alignment stats
read_statsDF <- alignStats(args)
read_statsDF <- cbind(read_statsDF[targets$FileName,], targets)
write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE, sep="\t")

## End(Not run)

#####
## Examples with \code{SYSargs2} object ##
#####
## Construct SYSargs2 object from CWL param, CWL input, and targets files
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF
names(WF); modules(WF); targets(WF)[1]; cmdlist(WF)[1:2]; output(WF)

## Not run:
## Execute SYSargs2 on single machine
WF <- runCommandline(args=WF)

## Execute SYSargs2 on multiple machines of a compute cluster.
file.copy(system.file("extdata", ".batchtools.conf.R", package="systemPipeR"), ".")
file.copy(system.file("extdata", "batchtools.slurm.tpl", package="systemPipeR"), ".")
resources <- list(walltime=120, ntasks=1, ncpus=4, memory=1024)
reg <- clusterRun(WF, FUN = runCommandline, more.args = list(args = WF, make_bam = TRUE), conffile=".batchtools")

## Monitor progress of submitted jobs
getStatus(reg=reg)

## Updates the path in the object \code{output(WF)}
WF <- output_update(WF, dir=FALSE, replace=TRUE, extension=c(".sam", ".bam"))

## Alignment stats
read_statsDF <- alignStats(WF)
read_statsDF <- cbind(read_statsDF[targets$FileName,], targets)
write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE, sep="\t")

## End(Not run)

```

`runDiff`*Differential abundance analysis for many range sets*

Description

Convenience wrapper function for `run_edgeR` and `run_DESeq2` to perform differential expression or abundance analysis iteratively for several count tables. The latter can be peak calling results for several samples or counts generated for different genomic feature types. The function also returns the filtering results and plots from `filterDEGs`.

Usage

```
runDiff(args, diffFct, targets, cmp, dbrfilter, ...)
```

Arguments

<code>args</code>	Object of class <code>SYSargs</code> or <code>SYSargs2</code> where <code>infile1(args)</code> specifies the paths to the tabular read count data files and outputs files to the result files.
<code>diffFct</code>	Defines which function should be used for the differential abundance analysis. Can be <code>diffFct=run_edgeR</code> or <code>diffFct=run_DESeq2</code> .
<code>targets</code>	<code>targets data.frame</code>
<code>cmp</code>	character matrix where comparisons are defined in two columns. This matrix should be generated with <code>readComp()</code> from the targets file. Values used for comparisons need to match those in the <code>Factor</code> column of the targets file.
<code>dbrfilter</code>	Named vector with filter cutoffs of format <code>c(Fold=2,FDR=1)</code> where <code>Fold</code> refers to the fold change cutoff (unlogged) and <code>FDR</code> to the p-value cutoff. Those values are passed on to the <code>filterDEGs</code> function.
<code>...</code>	Arguments to be passed on to the internally used <code>run_edgeR</code> or <code>run_DESeq2</code> function.

Value

Returns list containing the `filterDEGs` results for each count table. Each result set is a list with four components which are described under `?filterDEGs`. The result files contain the `edgeR` or `DESeq2` results from the comparisons specified under `cmp`. The base names of the result files are the same as the corresponding input files specified under `countfiles` and the value of extension appended.

Author(s)

Thomas Girke

See Also

`run_edgeR`, `run_DESeq2`, `filterDEGs`

Examples

```
## Paths to BAM files
param <- system.file("extdata", "bowtieSE.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args_bam <- systemArgs(sysma=param, mytargets=targets)
bfl <- BamFileList(outpaths(args_bam), yieldSize=50000, index=character())

## Not run:
## SYSargs with paths to range data and count files
args <- systemArgs(sysma="param/count_rangesets.param", mytargets="targets_macs.txt")

## Iterative read counting
countDFnames <- countRangeset(bfl, args, mode="Union", ignore.strand=TRUE)
writeTargetsout(x=args, file="targets_countDF.txt", overwrite=TRUE)

## Run differential abundance analysis
cmp <- readComp(file=args_bam, format="matrix")
args_diff <- systemArgs(sysma="param/rundiff.param", mytargets="targets_countDF.txt")
dbrlist <- runDiff(args, diffFct=run_edgeR, targets=targetsin(args_bam), cmp=cmp[[1]], independent=TRUE, dbrf=
writeTargetsout(x=args_diff, file="targets_rundiff.txt", overwrite=TRUE)

## End(Not run)
```

run_DESeq2

Runs DESeq2

Description

Convenience wrapper function to identify differentially expressed genes (DEGs) in batch mode with DESeq2 for any number of pairwise sample comparisons specified under the `cmp` argument. Users are strongly encouraged to consult the DESeq2 vignette for more detailed information on this topic and how to properly run DESeq2 on data sets with more complex experimental designs.

Usage

```
run_DESeq2(countDF, targets, cmp, independent = FALSE)
```

Arguments

<code>countDF</code>	date.frame containing raw read counts
<code>targets</code>	targets data.frame
<code>cmp</code>	character matrix where comparisons are defined in two columns. This matrix should be generated with the <code>readComp()</code> function from the targets file. Values used for comparisons need to match those in the Factor column of the targets file.
<code>independent</code>	If <code>independent=TRUE</code> then the <code>countDF</code> will be subsetted for each comparison. This behavior can be useful when working with samples from unrelated studies. For samples from the same or comparable studies, the setting <code>independent=FALSE</code> is usually preferred.

Value

data.frame containing DESeq2 results from all comparisons. Comparison labels are appended to column titles for tracking.

Author(s)

Thomas Girke

References

Please properly cite the DESeq2 papers when using this function: <http://www.bioconductor.org/packages/devel/bioc/html/>

See Also

run_edgeR, readComp and DESeq2 vignette

Examples

```
targetspath <- system.file("extdata", "targets.txt", package="systemPipeR")
targets <- read.delim(targetspath, comment="#")
cmp <- readComp(file=targetspath, format="matrix", delim="-")
countfile <- system.file("extdata", "countDFeByg.xls", package="systemPipeR")
countDF <- read.delim(countfile, row.names=1)
degseqDF <- run_DESeq2(countDF=countDF, targets=targets, cmp=cmp[[1]], independent=FALSE)
pval <- degseqDF[, grep("_FDR$", colnames(degseqDF)), drop=FALSE]
fold <- degseqDF[, grep("_logFC$", colnames(degseqDF)), drop=FALSE]
DEG_list <- filterDEGs(degDF=degseqDF, filter=c(Fold=2, FDR=10))
names(DEG_list)
DEG_list$Summary
```

run_edgeR

Runs edgeR

Description

Convenience wrapper function to identify differentially expressed genes (DEGs) in batch mode with the edgeR GML method for any number of pairwise sample comparisons specified under the `cmp` argument. Users are strongly encouraged to consult the edgeR vignette for more detailed information on this topic and how to properly run edgeR on data sets with more complex experimental designs.

Usage

```
run_edgeR(countDF, targets, cmp, independent = TRUE, paired = NULL, mdsplot = "")
```

Arguments

countDF	date.frame containing raw read counts
targets	targets data.frame
cmp	character matrix where comparisons are defined in two columns. This matrix should be generated with readComp() from the targets file. Values used for comparisons need to match those in the Factor column of the targets file.

independent	If independent=TRUE then the countDF will be subsetted for each comparison. This behavior can be useful when working with samples from unrelated studies. For samples from the same or comparable studies, the setting independent=FALSE is usually preferred.
paired	Defines pairs (character vector) for paired analysis. Default is unpaired (paired=NULL).
mdsplot	Directory where plotMDS should be written to. Default setting mdsplot="" will omit the plotting step.

Value

data.frame containing edgeR results from all comparisons. Comparison labels are appended to column titles for tracking.

Author(s)

Thomas Girke

References

Please properly cite the edgeR papers when using this function: <http://www.bioconductor.org/packages/devel/bioc/html/e>

See Also

run_DESeq2, readComp and edgeR vignette

Examples

```
targetspath <- system.file("extdata", "targets.txt", package="systemPipeR")
targets <- read.delim(targetspath, comment="#")
cmp <- readComp(file=targetspath, format="matrix", delim="-")
countfile <- system.file("extdata", "countDFeByg.xls", package="systemPipeR")
countDF <- read.delim(countfile, row.names=1)
edgeDF <- run_edgeR(countDF=countDF, targets=targets, cmp=cmp[[1]], independent=FALSE, mdsplot="")
pval <- edgeDF[, grep("_FDR$", colnames(edgeDF)), drop=FALSE]
fold <- edgeDF[, grep("_logFC$", colnames(edgeDF)), drop=FALSE]
DEG_list <- filterDEGs(degDF=edgeDF, filter=c(Fold=2, FDR=10))
names(DEG_list)
DEG_list$Summary
```

run_track

Keep track of the all SYSargs2 object

Description

Keep track of the all SYSargs2 object.

Usage

```
run_track(WF_ls)
```

Arguments

WF_ls list of SYSargs2 objects

Value

SYSargs2Pipe object

Author(s)

Daniela Cassol and Thomas Girke

See Also

showClass("SYSargs2Pipe") loadWorkflow renderWF

Examples

```
## Construct SYSargs2 object number 1
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF

## Construct SYSargs2 object number 2
targetsPE <- system.file("extdata", "targetsPE.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-pe", package="systemPipeR")
WF1 <- loadWorkflow(targets=targetsPE, wf_file="hisat2-mapping-pe.cwl",
                  input_file="hisat2-mapping-pe.yml", dir_path=dir_path)
WF1 <- renderWF(WF1, inputvars=c(FileName1="_FASTQ_PATH1_", FileName2="_FASTQ_PATH2_", SampleName="_SampleName_"))
WF1

## Keep track
WF_set <- run_track(WF_ls = c(WF1, WF))
WF_steps(WF_set)
track(WF_set)
summaryWF(WF_set)[1]
```

scaleRanges

Scale spliced ranges to genome coordinates

Description

Function to scale mappings of spliced features (query ranges) to their corresponding genome coordinates (subject ranges). The method accounts for introns in the subject ranges that are absent in the query ranges. A use case example are uORFs predicted in the 5' UTRs sequences using predORF. These query ranges are given relative to the 5' UTR sequence. The scaleRanges function will scale them to the corresponding genome coordinates. This way they can be used in RNA-Seq expression experiments like other gene ranges.

Usage

```
scaleRanges(subject, query, type = "custom", verbose = TRUE)
```

Arguments

subject	Genomic ranges provided as GRangesList object. Their name and length requirements are described under query.
query	Feature level ranges provided as GRangesList object. The names of the query ranges need to match the names of the GRangesList object assigned to the subject argument. In addition, the length of each query range cannot exceed the total length of the corresponding subject range set.
type	Feature name to use in type column of GRangesList result.
verbose	The setting verbose=FALSE suppresses all print messages.

Value

Object of class GRangesList

Author(s)

Thomas Girke

See Also

predORF

Examples

```
## Usage for simple example
subject <- GRanges(seqnames="Chr1", IRanges(c(5,15,30),c(10,25,40)), strand="+")
query <- GRanges(seqnames="myseq", IRanges(1, 9), strand="+")
scaleRanges(GRangesList(myid1=subject), GRangesList(myid1=query), type="test")

## Not run:
## Usage for more complex example
library(GenomicFeatures); library(systemPipeRdata)
gff <- system.file("extdata/annotation", "tair10.gff", package="systemPipeRdata")
txdb <- makeTxDbFromGFF(file=gff, format="gff3", organism="Arabidopsis")
futr <- fiveUTRsByTranscript(txdb, use.names=TRUE)
genome <- system.file("extdata/annotation", "tair10.fasta", package="systemPipeRdata")
dna <- extractTranscriptSeqs(FaFile(genome), futr)
uorf <- predORF(dna, n="all", mode="orf", longest_disjoint=TRUE, strand="sense")
grl_scaled <- scaleRanges(subject=futr, query=uorf, type="uORF", verbose=TRUE)
export.gff3(unlist(grl_scaled), "uorf.gff")

## End(Not run)
```

Description

The following `seeFastq` and `seeFastqPlot` functions generate and plot a series of useful quality statistics for a set of FASTQ files including per cycle quality box plots, base proportions, base-level quality trends, relative k-mer diversity, length and occurrence distribution of reads, number of reads above quality cutoffs and mean quality distribution. The functions allow processing of reads with variable length, but most plots are only meaningful if the read positions in the FASTQ file are aligned with the sequencing cycles. For instance, constant length clipping of the reads on either end or variable length clipping on the 3' end maintains this relationship, while variable length clipping on the 5' end without reversing the reads erases it.

The function `seeFastq` computes the summary stats and stores them in a relatively small list object that can be saved to disk with `save()` and reloaded with `load()` for later plotting. The argument `'klength'` specifies the k-mer length and `'batchsize'` the number of reads to random sample from each fastq file.

Usage

```
seeFastq(fastq, batchsize, klength = 8)
```

```
seeFastqPlot(fqlist, arrange = c(1, 2, 3, 4, 5, 8, 6, 7), ...)
```

Arguments

<code>fastq</code>	Named character vector containing paths to FASTQ file in the data fields and sample labels in the name slots.
<code>batchsize</code>	Number of reads to random sample from each FASTQ file that will be considered in the QC analysis. Smaller numbers reduce the memory footprint and compute time.
<code>klength</code>	Specifies the k-mer length in the plot for the relative k-mer diversity.
<code>fqlist</code>	list object returned by <code>seeFastq()</code> .
<code>arrange</code>	Integer vector from 1 to 7 specifying the row order of the QC plot. Dropping numbers eliminates the corresponding plots.
<code>...</code>	Additional plotting arguments to pass on to <code>seeFastqPlot()</code> .

Value

The function `seeFastq` returns the summary stats in a list containing all information required for the quality plots. The function `seeFastqPlot` plots the information generated by `seeFastq` using `ggplot2`.

Author(s)

Thomas Girke

Examples

```
## Not run:
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
args <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
args <- renderWF(args, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
fqlist <- seeFastq(fastq=infile1(args), batchsize=10000, klength=8)
```



```
pdf("fastqReport.pdf", height=18, width=4*length(fastq))
seeFastqPlot(fqlist)
dev.off()
```

```
## End(Not run)
```

subsetWF

Subsetting SYSargs2 class slots

Description

Return subsets of character for the input, output or the list of command-line for each workflow step.

Usage

```
subsetWF(args, slot, subset=NULL, index=NULL, delete=FALSE)
```

Arguments

args	object of class SYSargs2.
slot	three options available: type="input" returns input slot from SYSargs2 object; type="output" returns output slot from SYSargs2 object; and type="step" returns all the command-line for each workflow step from SYSargs2 object.
subset	name or numeric position of the values to be subsetting in the slot. If slot="input", the subset are the variables defined in the param.yml file. If slot="step", the subset is the command line defined on the SYSargs2 object for all the steps of the workflow. If slot="output", the subset is the path for the expected output files for all the steps in the workflow. Default is subset=NULL
index	A numeric index positions of the file in SYSargs2 object, slot output. It requires a subset to be defined. Default is index=NULL.
delete	allows to delete a subset of files in the case of slot="output". Default is delete=NULL.

Author(s)

Daniela Cassol and Thomas Girke

See Also

loadWorkflow renderWF

Examples

```
## Construct SYSargs2 object
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF
```

```
## Testing subset_wf function
input <- subsetWF(WF, slot="input", subset='FileName')
output <- subsetWF(WF, slot="output", subset=1, index=1)
step.cmd <- subsetWF(WF, slot="step", subset=1) ## subset all the HISAT2 commandline
# subsetWF(WF, slot="output", subset=1, index=1, delete=TRUE) ## in order to delete the subset files list
```

symLink2bam

Symbolic links for IGV

Description

Function for creating symbolic links to view BAM files in a genome browser such as IGV.

Usage

```
symLink2bam(sysargs, command="ln -s", htmldir, ext = c(".bam", ".bai"), urlbase, urlfile)
```

Arguments

sysargs	Object of class SYSargs or SYSargs2
command	Shell command, defaults to "ln -s"
htmldir	Path to HTML directory with http access.
ext	File name extensions to use for BAM and index files.
urlbase	The base URL structure to use in URL file.
urlfile	Name and path of URL file.

Value

symbolic links and url file

Author(s)

Thomas Girke

Examples

```
## Construct SYSargs object from param and targets files
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)

## Not run:
## Construct SYSargs2 object from cwl and targets files
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
args <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl", input_file="hisat2-mapping-se.yml", d
args <- renderWF(args, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))

## Create sym links and URL file for IGV
symLink2bam(sysargs=args, command="ln -s", htmldir=c("~/html/", "somedir/"), ext=c(".bam", ".bai"), urlbase=

## End(Not run)
```

 sysargs

SYSargs accessor methods

Description

Methods to access information from SYSargs object.

Usage

```
sysargs(x)
```

Arguments

x object of class SYSargs

Value

various outputs

Author(s)

Thomas Girke

Examples

```
## Construct SYSargs object from param and targets files
param <- system.file("extdata", "hisat2.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
args
names(args); modules(args); cores(args); outpaths(args); sysargs(args)
```

 SYSargs-class

Class "SYSargs"

Description

S4 class container for storing parameters of command-line- or R-based software. SYSargs instances are constructed by the `systemArgs` function from two simple tabular files: a `targets` file and a `param` file. The latter is optional for workflow steps lacking command-line software. Typically, a SYSargs instance stores all sample-level inputs as well as the paths to the corresponding outputs generated by command-line- or R-based software generating sample-level output files. Each sample level input/outfile operation uses its own SYSargs instance. The outpaths of SYSargs usually define the sample inputs for the next SYSargs instance. This connectivity is achieved by writing the outpaths with the `writeTargetsout` function to a new `targets` file that serves as input to the next `systemArgs` call. By chaining several SYSargs steps together one can construct complex workflows involving many sample-level input/output file operations with any combination of command-line or R-based software.

Objects from the Class

Objects can be created by calls of the form `new("SYSargs", ...)`.

Slots

targetsin: Object of class "data.frame" storing tabular data from targets input file
targetsout: Object of class "data.frame" storing tabular data from targets output file
targetsheader: Object of class "character" storing header/comment lines of targets file
modules: Object of class "character" storing software versions from module system
software: Object of class "character" name of executable of command-line software
cores: Object of class "numeric" number of CPU cores to use
other: Object of class "character" additional arguments
reference: Object of class "character" path to reference genome file
results: Object of class "character" path to results directory
infile1: Object of class "character" paths to first FASTQ file
infile2: Object of class "character" paths to second FASTQ file if data is PE
outfile1: Object of class "character" paths to output files generated by command-line software
sysargs: Object of class "character" full commands used to execute external software
outpaths: Object of class "character" paths to final outputs including postprocessing by Rsamtools

Methods

SampleName signature(x = "SYSargs"): extracts sample names
 [signature(x = "SYSargs"): subsetting of class with bracket operator
coerce signature(from = "list", to = "SYSargs"): as(list, "SYSargs")
cores signature(x = "SYSargs"): extracts data from cores slot
infile1 signature(x = "SYSargs"): extracts data from infile1 slot
infile2 signature(x = "SYSargs"): extracts data from infile2 slot
modules signature(x = "SYSargs"): extracts data from modules slot
names signature(x = "SYSargs"): extracts slot names
length signature(x = "SYSargs"): extracts number of samples
other signature(x = "SYSargs"): extracts data from other slot
outfile1 signature(x = "SYSargs"): extracts data from outfile1 slot
outpaths signature(x = "SYSargs"): extracts data from outpath slot
reference signature(x = "SYSargs"): extracts data from reference slot
results signature(x = "SYSargs"): extracts data from results slot
show signature(object = "SYSargs"): summary view of SYSargs objects
software signature(x = "SYSargs"): extracts data from software slot
targetsheader signature(x = "SYSargs"): extracts data from targetsheader slot
targetsin signature(x = "SYSargs"): extracts data from targetsin slot
targetsout signature(x = "SYSargs"): extracts data from targetsout slot

Author(s)

Thomas Girke

See Also

systemArgs and runCommandline

Examples

```

showClass("SYSargs")
## Construct SYSargs object from param and targets files
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
args
names(args); targetsin(args); targetsout(args); targetsheader(args);
software(args); modules(args); cores(args); outpaths(args)
sysargs(args); other(args); reference(args); results(args); infile1(args)
infile2(args); outfile1(args); SampleName(args)

## Return sample comparisons
readComp(args, format = "vector", delim = "-")

## The subsetting operator '[' allows to select specific samples
args[1:4]

## Not run:
## Execute SYSargs on single machine
runCommandline(args=args)

## Execute SYSargs on multiple machines
qsubargs <- getQsubargs(queue="batch", Nnodes="nodes=1", cores=cores(args), memory="mem=10gb", time="walltime")
qsubRun(appfct="runCommandline(args=args)", appargs=args, qsubargs=qsubargs, Nqsubs=1, submitdir="results",

## Write outpaths to new targets file for next SYSargs step
writeTargetsout(x=args, file="default")

## End(Not run)

```

SYSargs2-class

*Class "SYSargs2"***Description**

SYSargs2 class stores all the information and instructions needed for processing a set of input files with a specific command-line or a series of command-line within a workflow. The SYSargs2 S4 class object is created from the loadWorkflow and renderWF function, which populates all the command-line for each sample in each step of the particular workflow. Each sample level input/outfile operation uses its own SYSargs2 instance. The output of SYSargs2 define all the expected output files for each step in the workflow, which usually it is the sample input for the next step in an SYSargs2 instance. Between different instances, this connectivity is established by writing the subsetting output with the writeTargetsout function to a new targets file that serves as input to the next loadWorkflow and renderWF call. By chaining several SYSargs2 steps together

one can construct complex workflows involving many sample-level input/output file operations with any combination of command-line or R-based software.

Objects from the Class

Objects can be created by calls of the form `new("SYSargs2", ...)`.

Slots

targets: Object of class "list" storing data from each sample from targets file
targetsheader: Object of class "list" storing header/comment lines of targets file
modules: Object of class "list" storing software versions from module system
wf: Object of class "list" storing data from Workflow CWL param file
clt: Object of class "list" storing data from each CommandLineTool step in the Workflow or the single CommandLineTool CWL param file
yamlinput: Object of class "list" storing data from input file
cmdlist: Object of class "list" storing all command-line used to execute external software
input: Object of class "list" storing data from each target defined in inputvars
output: Object of class "list" paths to final outputs files
cwlfiles: Object of class "list" paths to input and CWL param files
inputvars: Object of class "list" storing data from each inputvars

Methods

[signature(x = "SYSargs2"): subsetting of class with bracket operator
[[signature(x = "SYSargs2", i = "ANY", j = "missing"): subsetting of class with bracket operator
[[<- signature(x = "SYSargs2"): replacement method for SYSargs2 class
\$ signature(x = "SYSargs2"): extracting slots elements by name
clt signature(x = "SYSargs2"): extracts data from clt slot
cmdlist signature(x = "SYSargs2"): extracts data from cmdlist slot
coerce signature(from = "list", to = "SYSargs2"): as(list, "SYSargs2")
cwlfiles signature(x = "SYSargs2"): extracts data from cwlfiles slot
infile1 signature(x = "SYSargs2"): extracting paths to first FASTQ file
infile2 signature(x = "SYSargs2"): extracting paths to second FASTQ file if data is PE
input signature(x = "SYSargs2"): extracts data from input slot
inputvars signature(x = "SYSargs2"): extracts data from inputvars slot
length signature(x = "SYSargs2"): extracts number of samples
modules signature(x = "SYSargs2"): extracts data from modules slot
names signature(x = "SYSargs2"): extracts slot names
output signature(x = "SYSargs2"): extracts data from cmdlist slot
show signature(object = "SYSargs2"): summary view of SYSargs2 objects
SYSargs2list signature(x = "SYSargs2"): Coerce back to list as(SYSargs2, "list")
targets signature(x = "SYSargs2"): extract data from targets slot
targetsheader signature(x = "SYSargs2"): extracts data from targetsheader slot
wf signature(x = "SYSargs2"): extracts data from wf slot
yamlinput signature(x = "SYSargs2"): extracts data from yamlinput slot

Author(s)

Daniela Cassol and Thomas Girke

See Also

loadWorkflow and renderWF and runCommandline and clusterRun

Examples

```
showClass("SYSargs2")

## Construct SYSargs2 object from CWL param, CWL input, and targets files
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF
names(WF); modules(WF); targets(WF)[1]; cmdlist(WF)[1:2]; output(WF)

## The subsetting operator '[' allows to select specific command-line
cmdlist(WF)[1:2]

## Not run:
## Execute SYSargs2 on single machine
WF <- runCommandline(args=WF)

## Execute SYSargs2 on multiple machines of a compute cluster. The following
## example uses the conf and template files for the Slurm scheduler. Please
## read the instructions on how to obtain the corresponding files for other schedulers.
file.copy(system.file("extdata", ".batchtools.conf.R", package="systemPipeR"), ".")
file.copy(system.file("extdata", "batchtools.slurm.tmpl", package="systemPipeR"), ".")
resources <- list(walltime=120, ntasks=1, ncpus=4, memory=1024)
reg <- clusterRun(args, FUN = runCommandline, conffile=".batchtools.conf.R", template="batchtools.slurm.tmpl")

## Monitor progress of submitted jobs
getStatus(reg=reg)

## Updates the path in the object \code{output(WF)}
WF <- output_update(WF, dir=FALSE, replace=TRUE, extension=c(".sam", ".bam"))

## Alignment stats
read_statsDF <- alignStats(WF)
read_statsDF <- cbind(read_statsDF[targets$FileName,], targets)
write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE, sep="\t")

## End(Not run)
```

SYSargs2list

SYSargs2 accessor methods

Description

Methods to access information from SYSargs2 object.

Usage

```
SYSargs2list(x)
```

Arguments

x object of class SYSargs2

Value

various outputs

Author(s)

Daniela Cassol and Thomas Girke

Examples

```
## Construct SYSargs2 object from CWL param, CWL input, and targets files
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF
names(WF); modules(WF); targets(WF)[1]; cmdlist(WF)[1:2]; output(WF)
```

SYSargs2Pipe-class *Class "SYSargs2Pipe"*

Description

SYSargs2Pipe class stores a list SYSargs2 objects. Each SYSargs2 objects stores all the information and instructions needed for processing a set of input files with a specific command-line or a series of command-line within a workflow.

Objects from the Class

Objects can be created by calls of the form `new("SYSargs2Pipe", ...)`.

Slots

WF_steps: Object of class "list" storing all the SYSargs2 objects

track: Object of class "list" storing all the output files from each SYSargs2 objects

summaryWF: Object of class "list" storing the summary of all the expected files exists and how many were missing for each SYSargs2 objects

Methods

```
[ signature(x = "SYSargs2Pipe", i = "ANY", j = "ANY", drop = "ANY"): subsetting of class with
  bracket operator
[[ signature(x = "SYSargs2Pipe", i = "ANY", j = "ANY"): subsetting of class with bracket oper-
  ator
[[<- signature(x = "SYSargs2Pipe", i = "ANY", j = "ANY", value = "ANY"): replacement method
  for SYSargs2 class
$ signature(x = "SYSargs2Pipe"): extracting slots elements by name
coerce signature(from = "list", to = "SYSargs2Pipe"): as(list, "SYSargs2Pipe")
coerce signature(from = "SYSargs2Pipe", to = "list"): as(SYSargs2Pipe, "list")
length signature(x = "SYSargs2Pipe"): extracts number of SYSargs2 objects
names signature(x = "SYSargs2Pipe"): extracts slot names
show signature(object = "SYSargs2Pipe"): summary view of SYSargs2 objects
summaryWF signature(x = "SYSargs2Pipe"): extract data from targets slot
SYSargs2Pipe_ls signature(x = "SYSargs2Pipe"): Coerce back to list as(SYSargs2Pipe, "list")
track signature(x = "SYSargs2Pipe"): extract data from track slot
WF_steps signature(x = "SYSargs2Pipe"): extract data from WF_steps slot
```

Author(s)

Daniela Cassol and Thomas Girke

See Also

loadWorkflow and renderWF and runCommandline and clusterRun

Examples

```
showClass("SYSargs2Pipe")

## Construct SYSargs2 object from CWL param, CWL input, and targets files
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF

## Keep track
WF_set <- run_track(WF_ls = c(WF))
WF_steps(WF_set)
track(WF_set)
summaryWF(WF_set)[1]
```

 SYSargs2Pipe_ls

SYSargs2Pipe accessor methods

Description

Methods to access information from SYSargs2Pipe object.

Usage

```
SYSargs2Pipe_ls(x)
```

Arguments

x object of class SYSargs2Pipe.

Value

various outputs

Author(s)

Daniela Cassol and Thomas Girke

Examples

```
## Construct SYSargs2 object number 1
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF
```

```
## Construct SYSargs2 object number 2
targetsPE <- system.file("extdata", "targetsPE.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-pe", package="systemPipeR")
WF1 <- loadWorkflow(targets=targetsPE, wf_file="hisat2-mapping-pe.cwl",
                  input_file="hisat2-mapping-pe.yml", dir_path=dir_path)
WF1 <- renderWF(WF1, inputvars=c(FileName1="_FASTQ_PATH1_", FileName2="_FASTQ_PATH2_", SampleName="_SampleName_"))
WF1
```

```
## Keep track
WF_set <- run_track(WF_ls = c(WF1, WF))
WF_steps(WF_set)
track(WF_set)
summaryWF(WF_set)[1]
```

systemArgs

Constructs SYSargs object from param and targets files

Description

Constructs SYSargs S4 class objects from two simple tabular files: a targets file and a param file. The latter is optional for workflow steps lacking command-line software. Typically, a SYSargs instance stores all sample-level inputs as well as the paths to the corresponding outputs generated by command-line- or R-based software generating sample-level output files. Each sample level input/outfile operation uses its own SYSargs instance. The outpaths of SYSargs usually define the sample inputs for the next SYSargs instance. This connectivity is established by writing the outpaths with the `writeTargetsout` function to a new targets file that serves as input to the next systemArgs call. By chaining several SYSargs steps together one can construct complex workflows involving many sample-level input/output file operations with any combination of command-line or R-based software.

Usage

```
systemArgs(sysma, mytargets, type = "SYSargs")
```

Arguments

sysma	path to 'param' file; file structure follows a simple name/value syntax that converted into JSON format; for details about the file structure see sample files provided by package. Assign NULL to run the pipeline without 'param' file. This can be useful for running partial workflows, e.g. with pregenerated BAM files.
mytargets	path to targets file
type	type="SYSargs" returns SYSargs, type="json" returns param file content in JSON format (requires rjson library)

Value

SYSargs object or character string in JSON format

Author(s)

Thomas Girke

See Also

```
showClass("SYSargs")
```

Examples

```
## Construct SYSargs object from param and targets files
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
args
names(args); modules(args); cores(args); outpaths(args); sysargs(args)

## Not run:
```

```

## Execute SYSargs on single machine
runCommandLine(args=args)

## Execute SYSargs on multiple machines of a compute cluster
resources <- list(walltime=120, ntasks=1, ncpus=cores(args), memory=1024)
reg <- clusterRun(args, conffile=".batchtools.conf.R", template="batchtools.slurm.tpl", Njobs=18, runid="01")

## Monitor progress of submitted jobs
getStatus(reg=reg)
file.exists(outpaths(args))
sapply(1:length(args), function(x) loadResult(reg, x)) # Works once all jobs have completed successfully.

## Alignment stats
read_statsDF <- alignStats(args)
write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE, sep="\t")

## Write outpaths to new targets file for next SYSargs step
writeTargetsout(x=args, file="default")

## End(Not run)

```

targets.as.df

Convert targets list to data.frame

Description

Convert list, which stores data from each target input file to data.frame object.

Usage

```
targets.as.df(x)
```

Arguments

x An object of the class "list" that stores data from each target input file, as targets(WF).

Value

data.frame containing all the input file information.

Author(s)

Daniela Cassol and Thomas Girke

Examples

```

targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
                  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
WF
targets.as.df(targets(WF))

```

variantReport

Generate Variant Report

Description

Functions for generating tabular variant reports including genomic context annotations and confidence statistics of variants. The annotations are obtained with utilities provided by the `VariantAnnotation` package and the variant statistics are retrieved from the input VCF files.

Usage

```
## Variant report
variantReport(args, txdb, fa, organism)

## Combine variant reports
combineVarReports(args, filtercol, ncol = 15)

## Create summary statistics of variants
varSummary(args)
```

Arguments

<code>args</code>	Object of class <code>SYSargs</code> or <code>SYSargs2</code> . The paths of the input VCF files are specified under <code>infile1(args)</code> and the paths of the output files under <code>outfile1(args)</code> or <code>output(args)</code> .
<code>txdb</code>	Annotation data stored as <code>TranscriptDb</code> object, which can be obtained from GFF/GTF files, BioMart, Bioc Annotation packages, UCSC, etc. For details see the vignette of the <code>GenomicFeatures</code> package. It is important to use here matching versions of the <code>txdb</code> and <code>fa</code> objects. The latter is the genome sequence used for read mapping and variant calling.
<code>fa</code>	<code>FaFile</code> object pointing to the sequence file of the corresponding reference genome stored in FASTA format or a <code>BSgenome</code> instance.
<code>organism</code>	Character vector specifying the organism name of the reference genome.
<code>filtercol</code>	Named character vector containing in the <code>name</code> field the column titles to filter on, and in the <code>data</code> field the corresponding values to include in the report. For instance, the setting <code>filtercol=c(Consequence="nonsynonymous")</code> will include only nonsynonymous variants listed in the <code>Consequence</code> column. To omit the filtering step, one can use the setting <code>filtercol="All"</code> .
<code>ncol</code>	Integer specifying the number of columns in the tabular input files. Default is set to 15.

Value

Tabular output files.

Author(s)

Thomas Girke

See Also

filterVars

Examples

```
## Alignment with BWA (sequentially on single machine)
param <- system.file("extdata", "bwa.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)
sysargs(args)[1]

## Not run:
system("bwa index -a bwtsv ./data/tair10.fasta")
bampaths <- runCommandline(args=args)

## Alignment with BWA (parallelized on compute cluster)
resources <- list(walltime=120, ntasks=1, ncpus=cores(args), memory=1024)
reg <- clusterRun(args, conffile=".batchtools.conf.R", template="batchtools.slurm.tmpl", Njobs=18, runid="01")

## Variant calling with GATK
## The following creates in the initial step a new targets file
## (targets_bam.txt). The first column of this file gives the paths to
## the BAM files created in the alignment step. The new targets file and the
## parameter file gatk.param are used to create a new SYSargs
## instance for running GATK. Since GATK involves many processing steps, it is
## executed by a bash script gatk_run.sh where the user can specify the
## detailed running parameters. All three files are expected to be located in the
## current working directory. Samples files for gatk.param and
## gatk_run.sh are available in the subdirectory ./inst/extdata/ of the
## source file of the systemPipeR package.
writeTargetsout(x=args, file="targets_bam.txt")
system("java -jar CreateSequenceDictionary.jar R=./data/tair10.fasta O=./data/tair10.dict")
# system("java -jar /opt/picard/1.81/CreateSequenceDictionary.jar R=./data/tair10.fasta O=./data/tair10.dict")
args <- systemArgs(sysma="gatk.param", mytargets="targets_bam.txt")
resources <- list(walltime=120, ntasks=1, ncpus=cores(args), memory=1024)
reg <- clusterRun(args, conffile=".batchtools.conf.R", template="batchtools.slurm.tmpl", Njobs=18, runid="01")
writeTargetsout(x=args, file="targets_gatk.txt")

## Variant calling with BCFtools
## The following runs the variant calling with BCFtools. This step requires in
## the current working directory the parameter file sambcf.param and the
## bash script sambcf_run.sh.
args <- systemArgs(sysma="sambcf.param", mytargets="targets_bam.txt")
resources <- list(walltime=120, ntasks=1, ncpus=cores(args), memory=1024)
reg <- clusterRun(args, conffile=".batchtools.conf.R", template="batchtools.slurm.tmpl", Njobs=18, runid="01")
writeTargetsout(x=args, file="targets_sambcf.txt")

## Filtering of VCF files generated by GATK
args <- systemArgs(sysma="filter_gatk.param", mytargets="targets_gatk.txt")
filter <- "totalDepth(vr) >= 2 & (altDepth(vr) / totalDepth(vr) >= 0.8) & rowSums(softFilterMatrix(vr))==4"
# filter <- "totalDepth(vr) >= 20 & (altDepth(vr) / totalDepth(vr) >= 0.8) & rowSums(softFilterMatrix(vr))==6"
filterVars(args, filter, varcaller="gatk", organism="A. thaliana")
writeTargetsout(x=args, file="targets_gatk_filtered.txt")

## Filtering of VCF files generated by BCFtools
args <- systemArgs(sysma="filter_sambcf.param", mytargets="targets_sambcf.txt")
```

```

filter <- "rowSums(vr) >= 2 & (rowSums(vr[,3:4])/rowSums(vr[,1:4]) >= 0.8)"
# filter <- "rowSums(vr) >= 20 & (rowSums(vr[,3:4])/rowSums(vr[,1:4]) >= 0.8)"
filterVars(args, filter, varcaller="bcftools", organism="A. thaliana")
writeTargetsout(x=args, file="targets_sambcf_filtered.txt")

## Annotate filtered variants from GATK
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_gatk_filtered.txt")
txdb <- loadDb("./data/tair10.sqlite")
fa <- FaFile(systemPipeR::reference(args))
variantReport(args=args, txdb=txdb, fa=fa, organism="A. thaliana")

## Annotate filtered variants from BCftools
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
txdb <- loadDb("./data/tair10.sqlite")
fa <- FaFile(systemPipeR::reference(args))
variantReport(args=args, txdb=txdb, fa=fa, organism="A. thaliana")

## Combine results from GATK
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_gatk_filtered.txt")
combineDF <- combineVarReports(args, filtercol=c(Consequence="nonsynonymous"))
write.table(combineDF, "./results/combineDF_nonsyn_gatk.xls", quote=FALSE, row.names=FALSE, sep="\t")

## Combine results from BCftools
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
combineDF <- combineVarReports(args, filtercol=c(Consequence="nonsynonymous"))
write.table(combineDF, "./results/combineDF_nonsyn_sambcf.xls", quote=FALSE, row.names=FALSE, sep="\t")

## Summary for GATK
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_gatk_filtered.txt")
write.table(varSummary(args), "./results/variantStats_gatk.xls", quote=FALSE, col.names = NA, sep="\t")

## Summary for BCftools
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
write.table(varSummary(args), "./results/variantStats_sambcf.xls", quote=FALSE, col.names = NA, sep="\t")

## Venn diagram of variants
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_gatk_filtered.txt")
varlist <- sapply(names(outpaths(args))[1:4], function(x) as.character(read.delim(outpaths(args)[x])$VARID))
vennset_gatk <- overLapper(varlist, type="vennsets")
args <- systemArgs(sysma="annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
varlist <- sapply(names(outpaths(args))[1:4], function(x) as.character(read.delim(outpaths(args)[x])$VARID))
vennset_bcf <- overLapper(varlist, type="vennsets")
vennPlot(list(vennset_gatk, vennset_bcf), mymain="", mysub="GATK: red; BCftools: blue", colmode=2, ccol=c("bl

## End(Not run)

```

vennPlot

Plot 2-5 way Venn diagrams

Description

Plotting function of 2-5 way Venn diagrams from 'VENNset' objects or count set vectors. A useful feature is the possibility to combine the counts from several Venn comparisons with the same number of label sets in a single Venn diagram.

Usage

```
vennPlot(x, mymain = "Venn Diagram", mysub = "default", setlabels = "default", yoffset = seq(0, 10, b
```

Arguments

<code>x</code>	VENNset or list of VENNset objects. Alternatively, a vector of Venn counts or a list of vectors of Venn counts can be provided as input. If several Venn comparisons are provided in a list then their results are combined in a single Venn diagram, where the count sets are organized above each other.
<code>mymain</code>	Main title of plot.
<code>mysub</code>	Subtitle of plot. Default <code>mysub="default"</code> reports the number of unique items in all sets, as well as the number of unique items in each individual set, respectively.
<code>setlabels</code>	The argument <code>setlabels</code> allows to provide a vector of custom sample labels. However, assigning the proper names in the name slots of the initial <code>setlist</code> is preferred for tracking purposes.
<code>yoffset</code>	The results from several Venn comparisons can be combined in a single Venn diagram by assigning to <code>x</code> a list with several VENNsets or count vectors. The positional offset of the count sets in the plot can be controlled with the <code>yoffset</code> argument. The argument setting <code>colmode</code> allows to assign different colors to each count set. For instance, with <code>colmode=2</code> one can assign to <code>ccol</code> a color vector or a list, such as <code>ccol=c("blue", "red")</code> or <code>ccol=list(1:8, 8:1)</code> .
<code>ccol</code>	Character or numeric vector to define colors of count values, e.g. <code>ccol=c("black", "black", "red")</code>
<code>colmode</code>	See argument <code>yoffset</code> .
<code>lcol</code>	Character or numeric vector to define colors of set labels, e.g. <code>lcol=c("red", "green")</code>
<code>lines</code>	Character or numeric vector to define colors of lines in plot.
<code>mylwd</code>	Defines line width of shapes used in plot.
<code>diacol</code>	See argument <code>type</code> .
<code>type</code>	Defines shapes used to plot 4-way Venn diagram. Default <code>type="ellipse"</code> uses ellipses. The setting <code>type="circle"</code> returns an incomplete 4-way Venn diagram as circles. This representation misses two overlap sectors, but is sometimes easier to navigate than the default ellipse version. The missing Venn intersects are reported below the Venn diagram. Their font color can be controlled with the argument <code>diacol</code> .
<code>ccex</code>	Controls font size for count values.
<code>lcex</code>	Controls font size for set labels.
<code>sepsplit</code>	Character used to separate sample labels in Venn counts.
<code>...</code>	Additional arguments to pass on.

Value

Venn diagram plot.

Note

The functions provided here are an extension of the Venn diagram resources on this site: <http://manuals.bioinformatics.ucr.edu/home/venn/Venn-Diagrams>

Author(s)

Thomas Girke

References

See examples in 'The Electronic Journal of Combinatorics': <http://www.combinatorics.org/files/Surveys/ds5/VennSymm>

See Also

overLapper, olBarplot

Examples

```
## Sample data
setlist <- list(A=sample(letters, 18), B=sample(letters, 16),
               C=sample(letters, 20), D=sample(letters, 22),
               E=sample(letters, 18), F=sample(letters, 22))

## 2-way Venn diagram
vennset <- overLapper(setlist[1:2], type="vennsets")
vennPlot(vennset)

## 3-way Venn diagram
vennset <- overLapper(setlist[1:3], type="vennsets")
vennPlot(vennset)

## 4-way Venn diagram
vennset <- overLapper(setlist[1:4], type="vennsets")
vennPlot(list(vennset, vennset))

## Pseudo 4-way Venn diagram with circles
vennPlot(vennset, type="circle")

## 5-way Venn diagram
vennset <- overLapper(setlist[1:5], type="vennsets")
vennPlot(vennset)

## Alternative Venn count input to vennPlot (not recommended!)
counts <- sapply(vennlist(vennset), length)
vennPlot(counts)

## 6-way Venn comparison as bar plot
vennset <- overLapper(setlist[1:6], type="vennsets")
olBarplot(vennset, mincount=1)

## Bar plot of standard intersect counts
intersect <- overLapper(setlist, type="intersects")
olBarplot(intersect, mincount=1)

## Accessor methods for VENNset/INTERSECTset objects
names(vennset)
names(intersect)
setlist(vennset)
intersectmatrix(vennset)
complexitylevels(vennset)
vennlist(vennset)
```

```

intersectlist(interset)

## Coerce VENNset/INTERSECTset object to list
as.list(vennset)
as.list(interset)

## Pairwise intersect matrix and heatmap
olMA <- sapply(names(setlist),
function(x) sapply(names(setlist),
function(y) sum(setlist[[x]] %in% setlist[[y]])))
olMA
heatmap(olMA, Rowv=NA, Colv=NA)

## Presence-absence matrices for large numbers of sample sets
interset <- overLapper(setlist=setlist, type="intersects", complexity=2)
(paMA <- intersectmatrix(interset))
heatmap(paMA, Rowv=NA, Colv=NA, col=c("white", "gray"))

```

VENNset-class

Class "VENNset"

Description

Container for storing Venn intersect results created by the `overLapper` function. The `setlist` slot stores the original label sets as vectors in a list; `intersectmatrix` organizes the label sets in a present-absent matrix; `complexitylevels` represents the number of comparisons considered for each comparison set as vector of integers; and `vennlist` contains the Venn intersect vectors.

Objects from the Class

Objects can be created by calls of the form `new("VENNset", ...)`.

Slots

setlist: Object of class "list": list of vectors
intersectmatrix: Object of class "matrix": binary matrix
complexitylevels: Object of class "integer": vector of integers
vennlist: Object of class "list": list of vectors

Methods

as.list signature(x = "VENNset"): coerces VENNset to list
coerce signature(from = "list", to = "VENNset"): as(list, "VENNset")
complexitylevels signature(x = "VENNset"): extracts data from complexitylevels slot
intersectmatrix signature(x = "VENNset"): extracts data from intersectmatrix slot
length signature(x = "VENNset"): returns number of original label sets
names signature(x = "VENNset"): extracts slot names
setlist signature(x = "VENNset"): extracts data from setlist slot
show signature(object = "VENNset"): summary view of VENNset objects
vennlist signature(x = "VENNset"): extracts data from vennset slot

file	Name and path of the output file. If set to "default" then the name of the output file will have the pattern 'targets_<software>.txt', where <software> will be what software(x) returns, when x is an object of class SYSargs. For an object of class SYSargs2, the output file will have the pattern 'targets_<software>_<step>.txt', where <software> will be the workflow name (cwlfiles(x)\$cwl) and <step> will be the step chosen in the argument step.
silent	If set to TRUE, all messages returned by the function will be suppressed.
overwrite	If set to TRUE, existing files of same name will be overwritten.
step	Name or numeric position of the step in the workflow to write the output files. The names can be check running names(x\$cwl).
new_col	A vector of character strings of the new column name to add to target file.
new_col_output_index	A vector of numeric index positions of the file in SYSargs2 class output. It requires new_col definition.
...	To pass on additional arguments.

Value

Writes tabular targets files containing the header/comment lines from targetsheader(x) and the columns from targetsout(x).

Author(s)

Daniela Cassol and Thomas Girke

See Also

writeTargetsRef

Examples

```
#####
## Examples with \code{SYSargs} object ##
#####
## Create SYSargs object
param <- system.file("extdata", "tophat.param", package="systemPipeR")
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
args <- systemArgs(sysma=param, mytargets=targets)

## Not run:
## Write targets out file
writeTargetsout(x=args, file="default")

## End(Not run)

#####
## Examples with \code{SYSargs2} object ##
#####
## Construct SYSargs2 object
targets <- system.file("extdata", "targets.txt", package="systemPipeR")
dir_path <- system.file("extdata/cwl/hisat2/hisat2-se", package="systemPipeR")
WF <- loadWorkflow(targets=targets, wf_file="hisat2-mapping-se.cwl",
  input_file="hisat2-mapping-se.yml", dir_path=dir_path)
WF <- renderWF(WF, inputvars=c(FileName="_FASTQ_PATH1_", SampleName="_SampleName_"))
```

```

WF

## Not run:
## Write targets out file
names(WF$c1t)
writeTargetsout(x=WF, file="default", step=1, new_col=c("sam_file"), new_col_output_index=c(1))

## End(Not run)

```

writeTargetsRef *Generate targets file with reference*

Description

Generates targets file with sample-wise reference as required for some NGS applications, such as ChIP-Seq containing input samples. The reference sample information needs to be provided in the input file in a column called SampleReference where the values reference the labels used in the SampleName column. Sample rows without reference assignments will be removed automatically.

Usage

```
writeTargetsRef(infile, outfile, silent = FALSE, overwrite = FALSE, ...)
```

Arguments

infile	Path to input targets file.
outfile	Path to output targets file.
silent	If set to TRUE, all messages returned by the function will be suppressed.
overwrite	If set to TRUE, existing files of same name will be overwritten.
...	To pass on additional arguments.

Value

Generates modified targets file with the paths to the reference samples in the second column named FileName2. Note, sample rows not assigned reference samples are removed automatically.

Author(s)

Thomas Girke

See Also

writeTargetsout, mergeBamByFactor

Examples

```

## Path to input targets file
targets <- system.file("extdata", "targets_chip.txt", package="systemPipeR")

## Not run:
## Write modified targets file with reference (e.g. input) sample
writeTargetsRef(infile=targets, outfile=~ /targets_refsampl.e.txt", silent=FALSE, overwrite=FALSE)

## End(Not run)

```

Index

*Topic **classes**

catDB-class, 4
INTERSECTset-class, 25
SYSargs-class, 59
SYSargs2-class, 61
SYSargs2Pipe-class, 64
VENNset-class, 74

*Topic **utilities**

alignStats, 3
catmap, 5
clusterRun, 6
countRangeset, 8
createWF, 10
featureCoverage, 11
featuretypeCounts, 14
filterDEGs, 16
filterVars, 17
genFeatures, 19
getQsubargs, 21
GOHyperGAll, 22
loadWorkflow, 26
mergeBamByFactor, 28
module, 29
moduleload, 30
olBarplot, 31
olRanges, 33
output_update, 33
overLapper, 35
plotfeatureCoverage, 37
plotfeaturetypeCounts, 39
predORF, 41
preprocessReads, 43
qsubRun, 44
readComp, 45
renderWF, 46
returnRPKM, 47
run_DESeq2, 51
run_edgeR, 52
run_track, 53
runCommandline, 48
runDiff, 50
scaleRanges, 54
seeFastq, 55

subsetWF, 57
symLink2bam, 58
sysargs, 59
SYSargs2list, 63
SYSargs2Pipe_ls, 66
systemArgs, 67
targets.as.df, 68
variantReport, 69
vennPlot, 71
writeTargetsout, 75
writeTargetsRef, 77

[, SYSargs, ANY, ANY, ANY-method
(SYSargs-class), 59
[, SYSargs2, ANY, ANY, ANY-method
(SYSargs2-class), 61
[, SYSargs2Pipe, ANY, ANY, ANY-method
(SYSargs2Pipe-class), 64
[[, SYSargs2, ANY, missing-method
(SYSargs2-class), 61
[[, SYSargs2Pipe, ANY, ANY-method
(SYSargs2Pipe-class), 64
[[<-, SYSargs2, ANY, ANY, ANY-method
(SYSargs2-class), 61
[[<-, SYSargs2Pipe, ANY, ANY, ANY-method
(SYSargs2Pipe-class), 64
\$, SYSargs2-method (SYSargs2-class), 61
\$, SYSargs2Pipe-method
(SYSargs2Pipe-class), 64

alignStats, 3
as.list, INTERSECTset-method
(INTERSECTset-class), 25
as.list, VENNset-method (VENNset-class),
74

catDB-class, 4
catlist (catmap), 5
catlist, catDB-method (catDB-class), 4
catlist-methods (catmap), 5
catmap, 5
catmap, catDB-method (catDB-class), 4
catmap-methods (catmap), 5
clt (SYSargs2list), 63
clt, SYSargs2-method (SYSargs2-class), 61

- clt-methods (SYSargs2list), 63
- clusterRun, 6
- cmdlist (SYSargs2list), 63
- cmdlist, SYSargs2-method (SYSargs2-class), 61
- cmdlist-methods (SYSargs2list), 63
- coerce, list, catDB-method (catDB-class), 4
- coerce, list, INTERSECTset-method (INTERSECTset-class), 25
- coerce, list, SYSargs-method (SYSargs-class), 59
- coerce, list, SYSargs2-method (SYSargs2-class), 61
- coerce, list, SYSargs2Pipe-method (SYSargs2Pipe-class), 64
- coerce, list, VENNset-method (VENNset-class), 74
- coerce, SYSargs2, list-method (SYSargs2-class), 61
- coerce, SYSargs2Pipe, list-method (SYSargs2Pipe-class), 64
- combineVarReports (variantReport), 69
- complexitylevels (overLapper), 35
- complexitylevels, INTERSECTset-method (INTERSECTset-class), 25
- complexitylevels, VENNset-method (VENNset-class), 74
- complexitylevels-methods (overLapper), 35
- cores (sysargs), 59
- cores, SYSargs-method (SYSargs-class), 59
- cores-methods (sysargs), 59
- countRangeset, 8
- createWF, 10
- cwlfiles (SYSargs2list), 63
- cwlfiles, SYSargs2-method (SYSargs2-class), 61
- cwlfiles-methods (SYSargs2list), 63

- featureCoverage, 11
- featuretypeCounts, 14
- filterDEGs, 16
- filterVars, 17

- genFeatures, 19
- getQsubargs, 21
- goBarplot (GOHyperGAll), 22
- GOCluster_Report (GOHyperGAll), 22
- GOHyperGAll, 22
- GOHyperGAll_Simplify (GOHyperGAll), 22
- GOHyperGAll_Subset (GOHyperGAll), 22

- idconv (catmap), 5
- idconv, catDB-method (catDB-class), 4
- idconv-methods (catmap), 5
- infile1 (sysargs), 59
- infile1, SYSargs-method (SYSargs-class), 59
- infile1, SYSargs2-method (SYSargs2-class), 61
- infile1-methods (sysargs), 59
- infile2 (sysargs), 59
- infile2, SYSargs-method (SYSargs-class), 59
- infile2, SYSargs2-method (SYSargs2-class), 61
- infile2-methods (sysargs), 59
- input (SYSargs2list), 63
- input, SYSargs2-method (SYSargs2-class), 61
- input-methods (SYSargs2list), 63
- inputvars (SYSargs2list), 63
- inputvars, SYSargs2-method (SYSargs2-class), 61
- inputvars-methods (SYSargs2list), 63
- intersectlist (overLapper), 35
- intersectlist, INTERSECTset-method (INTERSECTset-class), 25
- intersectlist-methods (overLapper), 35
- intersectmatrix (overLapper), 35
- intersectmatrix, INTERSECTset-method (INTERSECTset-class), 25
- intersectmatrix, VENNset-method (VENNset-class), 74
- intersectmatrix-methods (overLapper), 35
- INTERSECTset-class, 25

- length, INTERSECTset-method (INTERSECTset-class), 25
- length, SYSargs-method (SYSargs-class), 59
- length, SYSargs2-method (SYSargs2-class), 61
- length, SYSargs2Pipe-method (SYSargs2Pipe-class), 64
- length, VENNset-method (VENNset-class), 74
- loadWF (loadWorkflow), 26
- loadWorkflow, 26

- makeCATdb (GOHyperGAll), 22
- mergeBamByFactor, 28
- module, 29
- modulelist (moduleload), 30
- moduleload, 30

- modules (sysargs), 59
- modules, SYSargs-method (SYSargs-class), 59
- modules, SYSargs2-method (SYSargs2-class), 61
- modules-methods (sysargs), 59
- names, catDB-method (catDB-class), 4
- names, INTERSECTset-method (INTERSECTset-class), 25
- names, SYSargs-method (SYSargs-class), 59
- names, SYSargs2-method (SYSargs2-class), 61
- names, SYSargs2Pipe-method (SYSargs2Pipe-class), 64
- names, VENNset-method (VENNset-class), 74
- olBarplot, 31
- olRanges, 33
- other (sysargs), 59
- other, SYSargs-method (SYSargs-class), 59
- other-methods (sysargs), 59
- outfile1 (sysargs), 59
- outfile1, SYSargs-method (SYSargs-class), 59
- outfile1-methods (sysargs), 59
- outpaths (sysargs), 59
- outpaths, SYSargs-method (SYSargs-class), 59
- outpaths-methods (sysargs), 59
- output (SYSargs2list), 63
- output, SYSargs2-method (SYSargs2-class), 61
- output-methods (SYSargs2list), 63
- output_update, 33
- overLapper, 35
- plotfeatureCoverage, 37
- plotfeaturetypeCounts, 39
- predORF, 41
- preprocessReads, 43
- qsubRun, 44
- readComp, 45
- reference (sysargs), 59
- reference, SYSargs-method (SYSargs-class), 59
- reference-methods (sysargs), 59
- renderWF, 46
- results (sysargs), 59
- results, SYSargs-method (SYSargs-class), 59
- results-methods (sysargs), 59
- returnRPKM, 47
- run_DESeq2, 51
- run_edgeR, 52
- run_track, 53
- runCommandline, 48
- runDiff, 50
- SampleName (sysargs), 59
- SampleName, SYSargs-method (SYSargs-class), 59
- SampleName-methods (sysargs), 59
- scaleRanges, 54
- seeFastq, 55
- seeFastqPlot (seeFastq), 55
- setlist (overLapper), 35
- setlist, INTERSECTset-method (INTERSECTset-class), 25
- setlist, VENNset-method (VENNset-class), 74
- setlist-methods (overLapper), 35
- show, catDB-method (catDB-class), 4
- show, INTERSECTset-method (INTERSECTset-class), 25
- show, SYSargs-method (SYSargs-class), 59
- show, SYSargs2-method (SYSargs2-class), 61
- show, SYSargs2Pipe-method (SYSargs2Pipe-class), 64
- show, VENNset-method (VENNset-class), 74
- software (sysargs), 59
- software, SYSargs-method (SYSargs-class), 59
- software-methods (sysargs), 59
- subsetWF, 57
- summaryWF (SYSargs2Pipe_ls), 66
- summaryWF, SYSargs2Pipe-method (SYSargs2Pipe-class), 64
- summaryWF-methods (SYSargs2Pipe_ls), 66
- symLink2bam, 58
- sysargs, 59
- sysargs, SYSargs-method (SYSargs-class), 59
- SYSargs-class, 59
- sysargs-methods (sysargs), 59
- SYSargs2-class, 61
- SYSargs2list, 63
- SYSargs2list, SYSargs2-method (SYSargs2-class), 61
- SYSargs2list-method (SYSargs2list), 63
- SYSargs2Pipe-class, 64
- SYSargs2Pipe-method (SYSargs2Pipe_ls), 66

`SYSargs2Pipe_ls`, 66
`SYSargs2Pipe_ls`, `SYSargs2Pipe`-method
 (`SYSargs2Pipe`-class), 64
`systemArgs`, 67

`targets` (`SYSargs2list`), 63
`targets`, `SYSargs2`-method
 (`SYSargs2`-class), 61
`targets-methods` (`SYSargs2list`), 63
`targets.as.df`, 68
`targetsheader` (`sysargs`), 59
`targetsheader`, `SYSargs`-method
 (`SYSargs`-class), 59
`targetsheader`, `SYSargs2`-method
 (`SYSargs2`-class), 61
`targetsheader-methods` (`sysargs`), 59
`targetsin` (`sysargs`), 59
`targetsin`, `SYSargs`-method
 (`SYSargs`-class), 59
`targetsin-methods` (`sysargs`), 59
`targetnout` (`sysargs`), 59
`targetnout`, `SYSargs`-method
 (`SYSargs`-class), 59
`targetnout-methods` (`sysargs`), 59
`track` (`SYSargs2Pipe_ls`), 66
`track`, `SYSargs2Pipe`-method
 (`SYSargs2Pipe`-class), 64
`track-methods` (`SYSargs2Pipe_ls`), 66

`variantReport`, 69
`varSummary` (`variantReport`), 69
`vennlist` (`overLapper`), 35
`vennlist`, `VENNset`-method
 (`VENNset`-class), 74
`vennlist-methods` (`overLapper`), 35
`vennPlot`, 71
`VENNset`-class, 74

`wf` (`SYSargs2list`), 63
`wf`, `SYSargs2`-method (`SYSargs2`-class), 61
`wf-methods` (`SYSargs2list`), 63
`WF_steps` (`SYSargs2Pipe_ls`), 66
`WF_steps`, `SYSargs2Pipe`-method
 (`SYSargs2Pipe`-class), 64
`WF_steps-methods` (`SYSargs2Pipe_ls`), 66
`writeTargetsout`, 75
`writeTargetsRef`, 77

`yamlinput` (`SYSargs2list`), 63
`yamlinput`, `SYSargs2`-method
 (`SYSargs2`-class), 61
`yamlinput-methods` (`SYSargs2list`), 63