

# Package ‘dupRadar’

April 23, 2016

**Type** Package

**Title** Assessment of duplication rates in RNA-Seq datasets

**Version** 1.0.0

**Date** 2015-09-26

**Author** Sergi Sayols <sergisayolspuig@gmail.com>, Holger Klein  
<holger.klein@gmail.com>

**Maintainer** Sergi Sayols <sergisayolspuig@gmail.com>, Holger Klein  
<holger.klein@gmail.com>

**Description** Duplication rate quality control for RNA-Seq datasets.

**VignetteBuilder** knitr

**Suggests** BiocStyle, knitr, rmarkdown, AnnotationHub

**Depends** R (>= 3.2.0)

**Imports** Rsubread (>= 1.14.1)

**LazyData** true

**License** GPL-3

**biocViews** Technology, Sequencing, RNASeq, QualityControl

**NeedsCompilation** no

## R topics documented:

analyzeDuprates . . . . .	2
bamutilMarkDuplicates . . . . .	3
cumulativeDuprateBarplot . . . . .	4
dm . . . . .	4
dm.bad . . . . .	5
dupRadar . . . . .	5
dupRadar_examples . . . . .	6
duprateExpBoxplot . . . . .	6
duprateExpDensPlot . . . . .	7
duprateExpFit . . . . .	8

duprateExpIdentify . . . . .	8
duprateExpPlot . . . . .	9
expressionHist . . . . .	10
getBinDuplication . . . . .	11
getBinRpkMean . . . . .	11
getDupMatBin . . . . .	12
getDupMatStats . . . . .	12
getDynamicRange . . . . .	13
getRpkBinReadCountFraction . . . . .	14
getRpkCumulativeReadCountFraction . . . . .	14
markDuplicates . . . . .	15
picardMarkDuplicates . . . . .	16
readcountExpBoxplot . . . . .	16

<b>Index</b>	<b>18</b>
--------------	-----------

---

analyzeDuprates	<i>Read in a BAM file and count the tags falling on the features described in the GTF file</i>
-----------------	------------------------------------------------------------------------------------------------

---

## Description

analyzeDuprates returns a data.frame with tag counts

## Usage

```
analyzeDuprates(bam, gtf, stranded = 0, paired = FALSE, threads = 1,
  verbose = FALSE, ...)
```

## Arguments

bam	The bam file containing the duplicate-marked reads
gtf	The gtf file describing the features
stranded	Whether the reads are strand specific
paired	Paired end experiment?
threads	The number of threads to be used for counting
verbose	Whether to output Rsubread messages into the console
...	Other params sent to featureCounts

## Details

This function makes use of the Rsubread package to count tags on the GTF features in different scenarios. The scenarios are the 4 possible combinations of allowing multimappers (yes/no) and duplicate reads (yes/no).

**Value**

A data.frame with counts on features, with and without taking into account multimappers/duplicated reads

**Examples**

```
bam <- system.file("extdata",  
                  "wgEncodeCaltechRnaSeqGm12878R1x75dAlignsRep2V2_duprm.bam",  
                  package="dupRadar")  
gtf <- system.file("extdata", "genes.gtf", package="dupRadar")  
stranded <- 2 # '0' (unstranded), '1' (stranded) and '2' (reverse)  
paired <- FALSE  
threads <- 4  
  
# call the duplicate marker and analyze the reads  
dm <- analyzeDuprates(bam,gtf,stranded,paired,threads)
```

---

bamutilMarkDuplicates *Mark duplicates using bamutil*

---

**Description**

bamutilMarkDuplicates Mark duplicated reads from a BAM file by calling bamutil

**Usage**

```
bamutilMarkDuplicates(bam, out, path, verbose)
```

**Arguments**

bam	The bam file to mark duplicates from
out	Regular expression describing the transformation on the original filename to get the output filename. By default, a "_duprm" suffix is added before the bam extension
path	Path to the duplicate marker binaries
verbose	Redirect all the program output to the R console

**Details**

This function is supposed to be called through the markDuplicates wrapper

**Value**

The return code of the system call

---

`cumulativeDuprateBarplot`*Barplot showing the cumulative read counts fraction*

---

**Description**

`cumulativeDuprateBarplot` Barplot showing the cumulative read counts fraction

**Usage**

```
cumulativeDuprateBarplot(DupMat, stepSize = 0.05, ...)
```

**Arguments**

<code>DupMat</code>	The duplication matrix calculated by <code>analyzeDuprates</code>
<code>stepSize</code>	The size of the windows used for plotting
<code>...</code>	Other params sent to <code>barplot</code>

**Details**

This function makes a barplot showing the cumulative read counts fraction from the duplication matrix calculated by `analyzeDuprates`.

**Value**

nothing

**Examples**

```
# dm is a duplication matrix calculated by analyzeDuprates:
# R> dm <- analyzeDuprates(bamDuprm,gtf,stranded,paired,threads)
attach(dupRadar_examples)

# call the plot
cumulativeDuprateBarplot(DupMat=dm)
```

---

`dm`*Duplication matrix of a good RNASeq experiment*

---

**Description**

A dataset containing the duplication matrix of a good RNASeq experiment, in terms of duplicates. Comes from the GM12878 SR1x75 replicate 2 from Caltech (UCSC's table Browser name: wgEncodeCaltechRnaSeqGm12878R1x75dAlignsRep2V2)

**Usage**

```
data(dupRadar_examples)
```

**Format**

A data frame with 23228 rows and 14 variables

---

dm.bad	<i>Duplication matrix of a failed RNASeq experiment</i>
--------	---------------------------------------------------------

---

**Description**

A dataset containing the duplication matrix of a failed RNASeq experiment, containing unusual duplication rate. Comes from the HCT116 PE2x75 replicate 1 from Caltech (UCSC's table Browser name: wgEncodeCaltechRnaSeqHct116R2x75I1200AlignsRep1V2)

**Usage**

```
data(dupRadar_examples)
```

**Format**

A data frame with 23228 rows and 14 variables

---

dupRadar	<i>dupRadar.</i>
----------	------------------

---

**Description**

dupRadar.

---

dupRadar_examples	<i>Example data containing precomputed matrices for two RNASeq experiments</i>
-------------------	--------------------------------------------------------------------------------

---

### Description

Precomputed duplication matrices for two RNASeq experiments used as examples of a good and a failed (in terms of high redundancy of reads) experiments. The experiments come from the ENCODE project, as a source of a broad variety of protocols, library types and sequencing facilities.

### Usage

```
data(dupRadar_examples)
```

### Format

A list with two example duplication matrices

---

duprateExpBoxplot	<i>Duplication rate ~ total reads per kilobase (RPK) boxplot</i>
-------------------	------------------------------------------------------------------

---

### Description

duprateExpBoxplot Duplication rate ~ total reads per kilobase (RPK) boxplot

### Usage

```
duprateExpBoxplot(DupMat, stepSize = 0.05, ...)
```

### Arguments

DupMat	The duplication matrix calculated by analyzeDuprates
stepSize	Expression bin size for the boxplot
...	Other params sent to boxplot

### Details

This function makes a boxplot showing the distribution of per gene duplication rate versus the reads per kilobase (RPK) inside gene expression bins.

### Value

nothing

## Examples

```
# dm is a duplication matrix calculated by analyzeDuprates:
# R> dm <- analyzeDuprates(bamDuprm,gtf,stranded,paired,threads)
attach(dupRadar_examples)

# duprate boxplot
duprateExpBoxplot(DupMat=dm)
```

---

duprateExpDensPlot      *Duplication rate ~ total read count plot*

---

## Description

duprateExpDensPlot Duplication rate ~ total read count plot

## Usage

```
duprateExpDensPlot(DupMat, pal = c("cyan", "blue", "green", "yellow", "red"),
  tNoAlternative = TRUE, tRPKM = TRUE, tRPKMval = 0.5, tFit = TRUE,
  addLegend = TRUE, ...)
```

## Arguments

DupMat	The duplication matrix calculated by analyzeDuprates
pal	The color palette to use to display the density
tNoAlternative	Display threshold of 1000 reads per kilobase
tRPKM	Display threshold at a given RPKM level
tRPKMval	The given RPKM level
tFit	Whether to fit the model
addLegend	Whether to add a legend to the plot
...	Other parameters sent to plot()

## Details

This function makes a scatter plot showing the per gene duplication rate versus the total read count.

## Value

nothing

## Examples

```
# dm is a duplication matrix calculated by analyzeDuprates:
# R> dm <- analyzeDuprates(bamDuprm,gtf,stranded,paired,threads)
attach(dupRadar_examples)

# duprate plot
duprateExpDensPlot(DupMat=dm)
```

---

duprateExpFit      *Duplication rate ~ total read count fit model*

---

**Description**

duprateExpDensPlot Duplication rate ~ total read count fit model

**Usage**

```
duprateExpFit(DupMat)
```

**Arguments**

DupMat      The duplication matrix calculated by analyzeDuprates

**Details**

Fit a Generalized Linear Model using a logit function between the gene duplication rate and the total read count.

**Value**

The GLM and the coefficients of the fitted logit function

**Examples**

```
# dm is a duplication matrix calculated by analyzeDuprates:
# R> dm <- analyzeDuprates(bamDuprm,gtf,stranded,paired,threads)
attach(dupRadar_examples)

# duprate plot
duprateExpFit(DupMat=dm)
```

---

duprateExpIdentify      *Identify genes plotted by duprateExpPlot*

---

**Description**

duprateExpIdentify Identify genes plotted by duprateExpPlot

**Usage**

```
duprateExpIdentify(DupMat, idCol = "ID")
```



**Arguments**

DupMat            The duplication matrix calculated by analyzeDuprates  
 idCol            The column from the duplication matrix containing the labels

**Details**

This function makes a barplot showing the cumulative read counts fraction from the duplication matrix calculated by analyzeDuprates.

**Value**

The identified points. x and y values match the ones from duprateExpPlot

**Examples**

```
# dm is a duplication matrix calculated by analyzeDuprates:
# R> dm <- analyzeDuprates(bamDuprm,gtf,stranded,paired,threads)
attach(dupRadar_examples)

# call the plot and identify genes
duprateExpPlot(DupMat=dm)
duprateExpIdentify(DupMat=dm)
```

---

duprateExpPlot            *Duplication rate ~ total read count plot*

---

**Description**

duprateExpPlot Duplication rate ~ total read count plot

**Usage**

```
duprateExpPlot(DupMat, tNoAlternative = TRUE, tRPKM = TRUE,
  tRPKMval = 0.5, addLegend = TRUE, ...)
```

**Arguments**

DupMat            The duplication matrix calculated by analyzeDuprates  
 tNoAlternative    Display threshold of 1000 reads per kilobase  
 tRPKM            Display threshold at a given RPKM level  
 tRPKMval        The given RPKM level  
 addLegend        Whether to add a legend to the plot  
 ...              Other parameters sent to smoothScatter()

**Details**

This function makes a smooth scatter plot showing the per gene duplication rate versus the total read count.

**Value**

nothing

**Examples**

```
# dm is a duplication matrix calculated by analyzeDuprates:
# R> dm <- analyzeDuprates(bamDuprm,gtf,stranded,paired,threads)
attach(dupRadar_examples)

# duprate plot
duprateExpPlot(DupMat=dm)
```

---

expressionHist

*Draw histogram with the expression values*

---

**Description**

expressionHist Draw histogram with the expression values

**Usage**

```
expressionHist(DupMat, value = "RPK", ...)
```

**Arguments**

DupMat	The duplication matrix calculated by analyzeDuprates
value	The column from the duplication matrix containing the expression values
...	Other parameters sent to hist()

**Details**

This function draws a histogram of the expression values from the duplication matrix calculated by analyzeDuprates.

**Value**

nothing

**Examples**

```
# dm is a duplication matrix calculated by analyzeDuprates:
# R> dm <- analyzeDuprates(bamDuprm,gtf,stranded,paired,threads)
attach(dupRadar_examples)

# histogram of expression values for annotation
expressionHist(DupMat=dm)
```

---

getBinDuplication      *Helper function used in duprateExpBoxplot*

---

**Description**

getBinDuplication get duplication rate for a subset of the duplication matrix

**Usage**

```
getBinDuplication(p, stepSize, DupMat)
```

**Arguments**

p	The vector of bins
stepSize	The window size
DupMat	The duplication matrix calculated by analyzeDuprates

**Value**

The duplication rate per bin

---

getBinRpkMean      *Helper function used in duprateExpBoxplot*

---

**Description**

getBinRpkMean get mean duplication rate per bin

**Usage**

```
getBinRpkMean(p, stepSize, DupMat)
```

**Arguments**

p	The vector of bins
stepSize	The window size
DupMat	The duplication matrix calculated by analyzeDuprates

**Value**

The averaged RPK per bin

---

getDupMatBin	<i>Helper function used in getBinDuplication and getBinRpkMean</i>
--------------	--------------------------------------------------------------------

---

**Description**

getDupMatBin get a subset of the matrix for values in a specific bin defined by the upper bound p and stepSize

**Usage**

```
getDupMatBin(p, stepSize = 0.05, value = "allCounts", DupMat)
```

**Arguments**

p	The vector of bins
stepSize	The window size
value	The column to be subset
DupMat	The duplication matrix calculated by analyzeDuprates

**Value**

The subseted matrix

---

getDupMatStats	<i>Report duplication stats on regions</i>
----------------	--------------------------------------------

---

**Description**

getDupMatStats Report duplication stats based on the data calculated in the duplication matrix

**Usage**

```
getDupMatStats(DupMat)
```

**Arguments**

DupMat	The duplication matrix calculated by analyzeDuprates
--------	------------------------------------------------------

**Value**

A data.frame containing the stats about the number of genes covered (1+ tags) and the number of genes containing duplicates (1+)

**Examples**

```
# dm is a duplication matrix calculated by analyzeDuprates:
# R> dm <- analyzeDuprates(bamDuprm,gtf,stranded,paired,threads)
attach(dupRadar_examples)

# call the plot and identify genes
getDupMatStats(DupMat=dm)
```

---

getDynamicRange	<i>Dynamic range</i>
-----------------	----------------------

---

**Description**

getDynamicRange Calculate the dynamic range of the RNAseq experiment

**Usage**

```
getDynamicRange(dm)
```

**Arguments**

dm                    The duplication matrix calculated by analyzeDuprates

**Details**

This function calculates the dynamic range of the RNAseq experiment

**Value**

A list with 2 elements, containing the dynamic range counting all reads and the dynamic range after removing duplicates.

**Examples**

```
# dm is a duplication matrix calculated by analyzeDuprates:
# R> dm <- analyzeDuprates(bamDuprm,gtf,stranded,paired,threads)
attach(dupRadar_examples)

# calculate the dynamic range
getDynamicRange(dm)
```

---

`getRpkBinReadCountFraction`*Helper function used in readcountExpressionBoxplot*

---

**Description**

`readcountExpressionBoxplot` Calculates the fraction of total reads in a vector of bins

**Usage**

```
getRpkBinReadCountFraction(p, stepSize = stepSize, DupMat = DupMat)
```

**Arguments**

<code>p</code>	The vector of bins
<code>stepSize</code>	The window size
<code>DupMat</code>	The duplication matrix calculated by <code>analyzeDuprates</code>

**Value**

The fraction of total reads in a vector of bins

---

`getRpkCumulativeReadCountFraction`*Helper function used in readcountExpressionBoxplot*

---

**Description**

`getRpkCumulativeReadCountFraction` get the cumulative read count fraction

**Usage**

```
getRpkCumulativeReadCountFraction(p, DupMat = DupMat)
```

**Arguments**

<code>p</code>	The vector of bins
<code>DupMat</code>	The duplication matrix calculated by <code>analyzeDuprates</code>

**Value**

The cumulative read count fraction

---

markDuplicates	<i>Program dispatchers to mark duplicated reads from a BAM file</i>
----------------	---------------------------------------------------------------------

---

## Description

markDuplicates Mark duplicated reads from a BAM file by calling widely used tools.

## Usage

```
markDuplicates(dupremover = "bamutil", bam = NULL, out = gsub("\\.bam$",
  "_duprm.bam", bam), rminput = TRUE, path = ".", verbose = TRUE, ...)
```

## Arguments

dupremover	The tool to be called. Currently, "picard" and "bamutils" are supported
bam	The bam file to mark duplicates from
out	Regular expression describing the transformation on the original filename to get the output filename. By default, a "_duprm" suffix is added before the bam extension
rminput	Whether to keep the original, non duplicate-marked, bam file
path	Path to the duplicate marker binaries
verbose	Redirect all the program output to the R console
...	Other parameters sent to the caller function

## Details

This function works as a wrapper for several tools widely adopted to mark duplicated reads in a BAM file. Currently, it supports PICARD and BamUtils.

## Value

The output filename

## Examples

```
## Not run:
bam <- system.file("extdata", "sample1Aligned.out.bam", package="dupRadar")
gtf <- "genes.gtf"
stranded <- 2 # '0' (unstranded), '1' (stranded) and '2' (reverse)
paired <- FALSE
threads <- 4

# call the duplicate marker and analyze the reads
bamDuprm <- markDuplicates(dupremover="bamutil", bam,
  path="/opt/bamUtil-master/bin", rminput=FALSE)
dm <- analyzeDuprates(bamDuprm, gtf, stranded, paired, threads)

## End(Not run)
```

---

picardMarkDuplicates *Mark duplicates using Picard tools*

---

**Description**

picardMarkDuplicates Mark duplicated reads from a BAM file by calling picard tools

**Usage**

```
picardMarkDuplicates(bam, out, path, verbose, threads = 1, maxmem = "4g")
```

**Arguments**

bam	The bam file to mark duplicates from
out	Regular expression describing the transformation on the original filename to get the output filename. By default, a "_duprm" suffix is added before the bam extension
path	Path to the duplicate marker binaries
verbose	Redirect all the program output to the R console
threads	Number of threads to use
maxmem	Max memory assigned to the jvm

**Details**

This function is supposed to be called through the markDuplicates wrapper

**Value**

The return code of the system call

---

readcountExpBoxplot *Barplot of percentage of reads falling into expression bins*

---

**Description**

readcountExpBoxplot Barplot of percentage of reads falling into expression bins

**Usage**

```
readcountExpBoxplot(DupMat, stepSize = 0.05, ...)
```



**Arguments**

DupMat	The duplication matrix calculated by <code>analyzeDuprates</code>
stepSize	The number of bars to be shown
...	Other parameters sent to <code>barplot()</code>

**Details**

This function makes a barplot of percentage of reads falling into expression bins

**Value**

nothing Other parameters sent to `barplot()`

**Examples**

```
# dm is a duplication matrix calculated by analyzeDuprates:
# R> dm <- analyzeDuprates(bamDuprm,gtf,stranded,paired,threads)
attach(dupRadar_examples)

# barplot of percentage of reads falling into expression bins
readcountExpBoxplot(DupMat=dm)
```

# Index

## \*Topic **datasets**

- dm, [4](#)
- dm.bad, [5](#)
- dupRadar\_examples, [6](#)

[analyzeDuprates](#), [2](#)

[bamutilMarkDuplicates](#), [3](#)

[cumulativeDuprateBarplot](#), [4](#)

dm, [4](#)

dm.bad, [5](#)

dupRadar, [5](#)

dupRadar-package (dupRadar), [5](#)

dupRadar\_examples, [6](#)

duprateExpBoxplot, [6](#)

duprateExpDensPlot, [7](#)

duprateExpFit, [8](#)

duprateExpIdentify, [8](#)

duprateExpPlot, [9](#)

[expressionHist](#), [10](#)

[getBinDuplication](#), [11](#)

[getBinRpkMean](#), [11](#)

[getDupMatBin](#), [12](#)

[getDupMatStats](#), [12](#)

[getDynamicRange](#), [13](#)

[getRpkBinReadCountFraction](#), [14](#)

[getRpkCumulativeReadCountFraction](#), [14](#)

[markDuplicates](#), [15](#)

[picardMarkDuplicates](#), [16](#)

[readcountExpBoxplot](#), [16](#)