Package 'fastqcr'

January 3, 2019

Type Package

Title Quality Control of Sequencing Data

Version 0.1.2

Date 2018-12-23

Description 'FASTQC' is the most widely used tool for evaluating the quality of high throughput sequencing data.
It produces, for each sample, an html report and a compressed file containing the raw data.
If you have hundreds of samples, you are not going to open up each 'HTML' page.
You need some way of looking at these data in aggregate.
'fastqcr' Provides helper functions to easily parse, aggregate and analyze
'FastQC' reports for large numbers of samples. It provides a convenient solution for building
a 'Multi-QC' report, as well as, a 'one-sample' report with result interpretations.

License GPL-2

Encoding UTF-8

LazyData true

Depends R (>= 3.1.2)

Imports dplyr, grid, gridExtra, ggplot2, magrittr, readr (>= 1.3.0),
rmarkdown(>= 1.4), rvest, tibble, tidyr, scales, stats, utils,
xml2

Suggests knitr

URL http://www.sthda.com/english/rpkgs/fastqcr/

BugReports https://github.com/kassambara/fastqcr/issues

RoxygenNote 6.0.1

Collate 'utilities.R' 'fastqc.R' 'fastqc_install.R' 'qc_aggregate.R'
'qc_plot.R' 'qc_plot_collection.R' 'qc_problems.R' 'qc_read.R'
'qc_read_collection.R' 'qc_report.R' 'qc_unzip.R'

NeedsCompilation no

Author Alboukadel Kassambara [aut, cre]

Maintainer Alboukadel Kassambara <alboukadel.kassambara@gmail.com>

Repository CRAN

Date/Publication 2019-01-03 00:20:16 UTC

R topics documented:

fastqc *Run FastQC Tool*

Description

Run FastQC Tool

Usage

```
fastqc(fq.dir = getwd(), qc.dir = NULL, threads = 4,
  fastqc.path = "~/bin/FastQC/fastqc")
```

Arguments

fq.dir	path to the directory containing fastq files. Default is the current working directory.
qc.dir	path to the FastQC result directory. If NULL, a directory named fastqc_results is created in the current working directory.
threads	the number of threads to be used. Default is 4.
fastqc.path	path to fastqc program

Value

Create a directory containing the reports

Examples

```
## Not run:
# Run FastQC: generates a QC directory
fastqc(fq.dir)

## End(Not run)
```

`fastqc_install` *Install FastQC Tool*

Description

Install the FastQC Tool. To be used only on Unix system.

Usage

```
fastqc_install(url, dest.dir = "~/bin")
```

Arguments

`url`	url to download the latest version. If missing, the function will try to install the latest version from http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc.
`dest.dir`	destination directory to install the tool.

`qc_aggregate` *Aggregate FastQC Reports for Multiple Samples*

Description

Aggregate multiple FastQC reports into a data frame.

Usage

```
qc_aggregate(qc.dir = ".", progressbar = TRUE)

## S3 method for class 'qc_aggregate'
summary(object, ...)

qc_stats(object)
```

Arguments

`qc.dir`	path to the FastQC result directory to scan.
`progressbar`	logical value. If TRUE, shows a progress bar.
`object`	an object of class qc_aggregate.
`...`	other arguments.

Value

- **qc_aggregate()** returns an object of class qc_aggregate which is a (tibble) data frame with the following column names:
  - sample: sample names
  - module: fastqc modules
  - status: fastqc module status for each sample
  - tot.seq: total sequences (i.e.: the number of reads)
  - seq.length: sequence length
  - pct.gc: % of GC content
  - pct.dup: % of duplicate reads
- **summary**: Generates a summary of qc_aggregate. Returns a data frame with the following columns:
  - module: fastqc modules
  - nb_samples: the number of samples tested
  - nb_pass, nb_fail, nb_warn: the number of samples that passed, failed and warned, respectively.
  - failed, warned: the name of samples that failed and warned, respectively.
- **qc_stats**: returns a data frame containing general statistics of fastqc reports. columns are: sample, pct.dup, pct.gc, tot.seq and seq.length.

Functions

- `qc_aggregate`: Aggregate FastQC Reports for Multiple Samples
- `qc_stats`: Creates general statistics of fastqc reports.

Examples

```
# Demo QC dir
qc.dir <- system.file("fastqc_results", package = "fastqcr")
qc.dir

# List of files in the directory
list.files(qc.dir)

# Aggregate the report
qc <- qc_aggregate(qc.dir, progressbar = FALSE)
qc

# Generates a summary of qc_aggregate
summary(qc)

# General statistics of fastqc reports.
qc_stats(qc)
```

qc_fails *Inspect Problems in Aggregated FastQC Reports*

Description

Inspect problems in aggregated FastQC reports.

Usage

```
qc_fails(object, element = c("sample", "module"), compact = TRUE)

qc_warns(object, element = c("sample", "module"), compact = TRUE)

qc_problems(object, element = c("sample", "module"), name = NULL,
  status = c("FAIL", "WARN"), compact = TRUE)
```

Arguments

`object` an object of class qc_aggregate.

`element` character vector specifying which element to check for inspecting problems. Allowed values are one of c("sample", "module"). Default is "sample".

- If "sample", shows samples with more failed and/or warned modules
- If "module", shows moduled that failed and/or warned in the most samples

`compact` logical value. If TRUE, returns a compact output format; otherwise, returns a stretched format.

`name` character vector containing the names of modules and/or samples of interest. See qc_read for valid module names. If name specified, a stretched output format is returned by default unless you explicitly indicate compact = TRUE.

`status` character vector specifying the module status. Allowed values includes one or the combination of c("FAIL", "WARN"). If status = "FAIL", only modules with failed status are returned.

Value

- **qc_problems(), qc_fails(), qc_warns()**: returns a tibble (data frame) containing samples that had one or more modules with failure or warning. The format and the interpretation of the results depend on the argument ’element’, which value is one of c("sample", "module").
  - **If element = "sample" (default)**, results are samples with failed and/or warned modules. The results contain the following columns: sample (sample names), nb_problems (the number of modules with problems), module (the name of modules with problems).
  - **If element = "module"**, results are modules that failed and/or warned in the most samples. The results contain the following columns: module (the name of module with problems), nb_problems (the number of samples with problems), sample (the name of samples with problems)

Functions

- `qc_fails`: Displays which samples had one or more failed modules. Use qc_fails(qc, "module") to see which modules failed in the most samples.
- `qc_warns`: Displays which samples had one or more warned modules. Use qc_warns(qc, "module") to see which modules warned in the most samples.
- `qc_problems`: Union of `qc_fails()` and `qc_warns()`. Display which samples or modules that failed or warned.

Examples

```
# Demo QC dir
qc.dir <- system.file("fastqc_results", package = "fastqcr")
qc.dir
# List of files in the directory
list.files(qc.dir)

# Aggregate the report
qc <- qc_aggregate(qc.dir, progressbar = FALSE)

# Display samples with failed modules
qc_fails(qc)
qc_fails(qc, compact = FALSE)

# Display samples with warned modules
qc_warns(qc)

# Module failed in the most samples
qc_fails(qc, "module")
qc_fails(qc, "module", compact = FALSE)

# Specify a module of interest
qc_problems(qc, "module",  name = "Per sequence GC content")
```

`qc_plot` *Plot FastQC Results*

Description

Plot FastQC data

Usage

```
qc_plot(qc, modules = "all")

## S3 method for class 'qctable'
print(x, ...)
```

Arguments

`qc` An object of class qc_read or a path to the sample zipped fastqc result file.

`modules` Character vector containing the names of fastqc modules for which you want to import the data. Default is all. Allowed values include one or the combination of:

- "Summary",
- "Basic Statistics",
- "Per base sequence quality",
- "Per sequence quality scores",
- "Per base sequence content",
- "Per sequence GC content",
- "Per base N content",
- "Sequence Length Distribution",
- "Sequence Duplication Levels",
- "Overrepresented sequences",
- "Adapter Content",
- "Kmer Content"

Partial match of module names allowed. For example, you can use modules = "GC content", instead of the full names modules = "Per sequence GC content".

`x` an object of class qctable.

`...` other arguments.

Value

Returns a list of ggplots containing the plot for specified modules..

Examples

```
# Demo file
qc.file <- system.file("fastqc_results", "S1_fastqc.zip",  package = "fastqcr")
qc.file
# Read all modules
qc <- qc_read(qc.file)

# Plot per sequence GC content
qc_plot(qc, "Per sequence GC content")

# Per base sequence quality
qc_plot(qc, "Per base sequence quality")

# Per sequence quality scores
qc_plot(qc, "Per sequence quality scores")

# Per base sequence content
qc_plot(qc, "Per base sequence content")

# Sequence duplication levels
```

```
qc_plot(qc, "Sequence duplication levels")
```

qc_plot_collection *Plot FastQC Results of multiple samples*

Description

Plot FastQC data of multiple samples

Usage

```
qc_plot_collection(qc, modules = "all")
```

Arguments

qc	An object of class qc_read_collection or a path to the sample zipped fastqc result files.
modules	Character vector containing the names of fastqc modules for which you want to import the data. Default is all. Allowed values include one or the combination of: • "Summary", • "Basic Statistics", • "Per base sequence quality", • "Per sequence quality scores", • "Per base sequence content", • "Per sequence GC content", • "Per base N content", • "Sequence Length Distribution", • "Sequence Duplication Levels", • "Overrepresented sequences", • "Adapter Content", • "Kmer Content" Partial match of module names allowed. For example, you can use modules = "GC content", instead of the full names modules = "Per sequence GC content".

Value

Returns a list of ggplots containing the plot for specified modules..

Author(s)

Mahmoud Ahmed, <mahmoud.s.fahmy@students.kasralainy.edu.eg>

Examples

```
qc.dir <- system.file("fastqc_results", package = "fastqcr")
qc.files <- list.files(qc.dir, full.names = TRUE)

# read all modules in all files
qc <- qc_read_collection(qc.files, sample_names = paste('S', 1:5, sep = ''))

# Plot per sequence GC content
qc_plot_collection(qc, "Per sequence GC content")

# Per base sequence quality
qc_plot_collection(qc, "Per base sequence quality")

# Per sequence quality scores
qc_plot_collection(qc, "Per sequence quality scores")

# Per base sequence content
qc_plot_collection(qc, "Per base sequence content")

# Sequence duplication levels
qc_plot_collection(qc, "Sequence duplication levels")
```

qc_read *Read FastQC Data*

Description

Read FastQC data into R.

Usage

```
qc_read(file, modules = "all", verbose = TRUE)
```

Arguments

`file` Path to the file to be imported. Can be the path to either :

- the fastqc zipped file (e.g.: 'path/to/samplename_fastqc.zip'). No need to unzip,
- or the unzipped folder name (e.g.: 'path/to/samplename_fastqc'),
- or the sample name (e.g.: 'path/to/samplename')
- or the fastqc_data.txt file,

`modules` Character vector containing the names of FastQC modules for which you want to import/inspect the data. Default is all. Allowed values include one or the combination of:

- "Summary",
- "Basic Statistics",

- "Per base sequence quality",
- "Per tile sequence quality",
- "Per sequence quality scores",
- "Per base sequence content",
- "Per sequence GC content",
- "Per base N content",
- "Sequence Length Distribution",
- "Sequence Duplication Levels",
- "Overrepresented sequences",
- "Adapter Content",
- "Kmer Content"

Partial match of module names allowed. For example, you can use modules = "GC content", instead of the full names modules = "Per sequence GC content".

`verbose` logical value. If TRUE, print filename when reading.

Value

Returns a list of tibbles containing the data for specified modules.

Examples

```
# Demo file
qc.file <- system.file("fastqc_results", "S1_fastqc.zip",  package = "fastqcr")
qc.file
# Read all modules
qc_read(qc.file)

# Read a specified module
qc_read(qc.file,"Per base sequence quality")
```

`qc_read_collection` *Read a collection of FastQC data files*

Description

A wrapper function around qc_read to read multiple FastQC data files at once.

Usage

```
qc_read_collection(files, sample_names, modules = "all", verbose = TRUE)
```

Arguments

`files`	A `character` vector of paths to the files to be imported.
`sample_names`	A `character` vector of length equals that of the first argument `files`
`modules`	Character vector containing the names of FastQC modules for which you want to import/inspect the data. Default is all. Allowed values include one or the combination of: • "Summary", • "Basic Statistics", • "Per base sequence quality", • "Per tile sequence quality", • "Per sequence quality scores", • "Per base sequence content", • "Per sequence GC content", • "Per base N content", • "Sequence Length Distribution", • "Sequence Duplication Levels", • "Overrepresented sequences", • "Adapter Content", • "Kmer Content" Partial match of module names allowed. For example, you can use modules = "GC content", instead of the full names modules = "Per sequence GC content".
`verbose`	logical value. If TRUE, print filename when reading.

Value

A `list` of `tibbles` containing the data of specified modules form each file.

Author(s)

Mahmoud Ahmed, <`mahmoud.s.fahmy@students.kasralainy.edu.eg`>

Examples

```
# extract paths to the demo files
qc.dir <- system.file("fastqc_results", package = "fastqcr")
qc.files <- list.files(qc.dir, full.names = TRUE)

# read all modules in all files
qc <- qc_read_collection(qc.files, sample_names = paste('S', 1:5, sep = ''))


# read a specified module in all files
qc <- qc_read_collection(qc.files,
    sample_names = paste('S', 1:5, sep = ''),
    modules = "Per base sequence quality")
```

qc_report *Build a QC Report*

Description

Create an HTML file containing FastQC reports of one or multiple files. Inputs can be either a directory containing multiple FastQC reports or a single sample FastQC report.

Usage

```
qc_report(qc.path, result.file, experiment = NULL, interpret = FALSE,
  template = NULL, preview = TRUE)
```

Arguments

qc.path	path to the FastQC reports. Allowed values include: • A path to a directory containing multiple zipped FastQC reports, • Or a single sample zipped FastQC report. Partial match is allowed for sample name.
result.file	path to the result file prefix (e.g., path/to/qc-result). Don't add the file extension.
experiment	text specifying a short description of the experiment. For example experiment = "RNA sequencing of colon cancer cell lines".
interpret	logical value. If TRUE, adds the interpretation of each module.
template	a character vector specifying the path to an Rmd template. file.
preview	logical value. If TRUE, shows a preview of the report.

Examples

```
## Not run:
# Demo QC Directory
qc.path <- system.file("fastqc_results", package = "fastqcr")
qc.path

# List of files in the directory
list.files(qc.path)

# Multi QC report
qc_report(qc.path, result.file = "~/Desktop/result")

# QC Report of one sample with plot interpretation
 qc.file <- system.file("fastqc_results", "S1_fastqc.zip", package = "fastqcr")
 qc_report(qc.file, result.file = "~/Desktop/result",
   interpret = TRUE)

## End(Not run)
```

qc_unzip *Unzip Files in the FastQC Result Directory*

Description

Unzip all files in the FastQC result directory. Default is the current working directory.

Usage

```
qc_unzip(qc.dir = ".", rm.zip = TRUE)
```

Arguments

`qc.dir` Path to the FastQC result directory.

`rm.zip` logical. If TRUE, remove zipped files after extraction. Default is TRUE.

Examples

```
## Not run:
qc_unzip("FASTQC")

## End(Not run)
```

Index