Package 'qckit'

February 27, 2018

Type Package

Title What the Package Does (Title Case)

Version 0.1.0

Author Who wrote it

Maintainer The package maintainer <yourself@somewhere.net>

Description More about what it does (maybe more than one line)
Use four spaces when indenting paragraphs within the Description.

License What license is it under?

Encoding UTF-8

LazyData false

Suggests testthat

RoxygenNote 6.0.1.9000

LinkingTo Rcpp

Imports magrittr,
ggplot2,
dplyr,
seqTools,
RSQLite,
zlibbioc,
Rcpp

R topics documented:

`basic_stat` *Generate the data frame that includes percentiles of quality score per position*

Description

Generate the data frame that includes percentiles of quality score per position

Usage

```
basic_stat(infile)
```

Arguments

`infile`	the object that is the dataframe of the mean, median and quantiles of the FASTQ file from basic statistics function

Value

boxplot of per position quality score distribution

`calc_over_rep_seq` *calculate Over Rep seqs*

Description

Calculate sequece counts for each unique sequence and create a table with unique sequences and corresponding counts

Usage

```
calc_over_rep_seq(infile, out_prefix, min_size = 5L, buffer_size = 1000000L)
```

Arguments

`infile`	A string giving the path for the fastqfile
`out_prefix`	A string giving the prefix to be used for outputs
`min_size`	An int for thhresholding over representation
`buffer_size`	An int for the number of lines to keep in memory

dimensions — *Extract the dimensions for Fastq file*

Description

ncolumnuse seqTool to extract the dimensions of a Fastq G zipped file

Usage

```
dimensions(fseq, selection)
```

Arguments

fseq	an object that is the read result of the seq.read function
selection	"reads' for number of reads/rows, 'positions' for number of positions/columns

Value

a numeric value of the number of reads or the number of positions

GC_content — *Extract GC content separately and calculate GC content percentage for each sequence read*

Description

Extract GC content separately and calculate GC content percentage for each sequence read

Usage

```
GC_content(infile)
```

Arguments

infile	the object that is the path to the FASTQ file

Value

plot of GC content

GC_content_plot *Generate GC content plot from the GC content*

Description

Generate GC content plot from the GC content

Usage

```
GC_content_plot(nc, gc_df, writefile = FALSE, prefix)
```

Arguments

nc	the object that is the number of positions of the FASTQ files
gc_df	the object that is the GC content vectors generated from GC content function
writefile	the object indicating intent to save the plot as pdf file, set default as FALSE
prefix	the prefix for the output file of the plot

Value

a ggplot of the GC content acorss all positions

gc_per_read *calculate GC percent per read*

Description

Calculate GC nucleotide sequence content per read of the FASTQ gzipped file

Usage

```
gc_per_read(infile)
```

Arguments

infile	A string giving the path for the fastqfile

kmer — *Extract kmers and kmer counts from FASTQ file to a data frame*

Description

Extract kmers and kmer counts from FASTQ file to a data frame

Usage

```
kmer(name, kcount, writefile = FALSE)
```

Arguments

name	the object that is the path to gzipped FASTQ file
kcount	the object that is the length of kmer that is in interest
writefile	the boolean object that asks whether to write output as csv file

Value

data frame of kmer and corresponding kmer count of the length of choice

overrepresented_sequence — *Sort all sequences per read by count along with a density plot of all counts with top 5 repreated sequences marked*

Description

Sort all sequences per read by count along with a density plot of all counts with top 5 repreated sequences marked

Usage

```
overrepresented_sequence(infile, nr, prefix)
```

Arguments

infile	the object that is the path to gzippped FASTQ file
nr	the number of reads of the FASTQ file
prefix	the prefix to name the output file

Value

table of sequnces sortted by count

density plot of sequence length with top 5 marked by rugs, saved as PDF file

overrep_kmer *Generate overrepresented kmers from all kmer counts results*

Description

Generate overrepresented kmers from all kmer counts results

Usage

```
overrep_kmer(path, k, nc, nr)
```

Arguments

path	the path to the gz file
k	the length of the sequence looking for
nc	number of positions
nr	number of reads

Value

the index of reads that has overrepresented kmers

overrep_plot *Plot the top 5 seqeunces*

Description

Plot the top 5 seqeunces

Usage

```
overrep_plot(overrep_order, prefix)
```

Arguments

overrep_order	the table that sorts the sequence content and corresponding counts in descending order
prefix	the prefix to the file saved

Value

plot of the top 5 overrepresented sequences

plotSeqContent *Plot the per position nucleotide content* plotSeqContent

Description

Plot the per position nucleotide content plotSeqContent

Usage

```
plotSeqContent(fseq, nr, nc, writefile = FALSE, prefix)
```

Arguments

nr	the number of reads of the FASTQ file, acquired through previous functions
nc	the number of positions of the FASTQ file, acquired through previous functions
writefile	the boolean object to write the plot as PDF file, default is FALSE
prefix	the prefix to add to the file name
name	the object that is the path to the gzipped FASTQ file

Value

ggplot line plot of all nucleotide content inclding A, T, G, C and N

plot_perseq_quality *plot the mean quality score per read in histograms* plot_perseq_quality

Description

plot the mean quality score per read in histograms plot_perseq_quality

Usage

```
plot_perseq_quality(infile, writefile = FALSE, prefix)
```

Arguments

infile	the object that is the path to the file that
writefile	the object indicating intent to save the plot as pdf file, set default as FALSE
prefix	the prefix for the output file of the plot

Value

plot of mean quality score per read

plot_quality_score *Generate a boxplot of the per posistion quality score from basic statistics results*

Description

Generate a boxplot of the per posistion quality score from basic statistics results

Usage

```
plot_quality_score(basic_statistics, writefile = FALSE, prefix)
```

Arguments

writefile	the object indicating intent to save the plot as pdf file, set default as FALSE
prefix	the prefix for the output file of the plot
basic_stat	the object that is the dataframe of the mean, median and quantiles of the FASTQ file from basic statistics function

Value

boxplot of per position quality score distribution

plot_sequence_length *extract the sequence length per read and plot corresponding bar plot*

Description

extract the sequence length per read and plot corresponding bar plot

Usage

```
plot_sequence_length(fseq, writefile = FALSE, prefix)
```

Arguments

fseq	the object that is the seqTools processed result
writefile	the boolean object to write the plot as PDF file, default is FALSE
prefix	the prefix to add to the file name

Value

the plot of the sequence distribution among all reads

Author(s)

Wenyue Xing, <wenyue_xing@brown.edu>

process_fastq *calculate Over Rep seqs*

Description

calculate Over Rep seqs

Usage

```
process_fastq(infile, out_prefix, buffer_size)
```

Arguments

infile	A string giving the path for the fastqfile
out_prefix	A string giving the prefix to be used for outputs
buffer_size	An int for the number of lines to keep in memory

qual_score_per_read *calculate mean quality per read*

Description

Calculate the mean quality score per read of the FASTQ gzipped file

Usage

```
qual_score_per_read(infile)
```

Arguments

infile	A string giving the path for the fastqfile

sequence_content *Extract nucleoctide sequence content per position from fastq file*

Description

Extract nucleoctide sequence content per position from fastq file

Usage

```
sequence_content(fseq, content)
```

Arguments

fseq	an object that is the read result from seq.read function
content	an object of string type that specifies the content in question, "A","T","G","C","N"(either capital or lower case)

Value

the per position

Author(s)

Wenyue Xing, <wenyue_xing@brown.edu>

Index