ReQON

March 24, 2012

`FWSEplot` *Plot reported vs. empirical quality.*

Description

Plots reported vs. empirical quality scores. Also calculates and outputs Frequency-Weighted Squared Error (FWSE) and reports FWSE on the plot.

Usage

```
FWSEplot(ErrRates, QualFreq, FWSE_out = TRUE, col = "blue", min_freq = 0.001,
   lim = c(0, length(QualFreq) - 1), xlabel = "Reported Quality",
   ylabel = "Empirical Quality", main_title = "Reported vs. Empirical Quality")
```

Arguments

`ErrRates`	vector of empirical error rates on the Phred scale.
`QualFreq`	vector of relative frequencies of quality scores.
`FWSE_out`	option to output FWSE and report FWSE on the plot. Default = TRUE.
`col`	color of plotted points. Default = "blue".
`min_freq`	Any quality scores with relative frequency less than this vlue will be plotted with a solid point. Default = 0.001.
`lim`	common axis limits for both the x-axis and y-axis. Default = c(0, length(QualFreq) - 1).
`xlabel`	x-axis label. Default = "Reported Quality".
`ylabel`	y-axis label. Default = "Empirical Quality".
`main_title`	title. Default = "Reported vs. Empirical Quality".

Details

FWSEplot plots the reported quality score against the empirical quality score and reports FWSE. If the quality scores accurately reflect the probability of a sequencing error, then the points should fall close to the 45-degree line and FWSE should be close to zero. If the input vectors are $ErrRatesBefore and $QualFreqBefore from the ReQON output, this function will create the bottom left diagnostic plot that is output from ReQON. Similarly, if the input vectors are $ErrRatesAfter and $QualFreqAfter from ReQON output, then the bottom right diagnostic plot is created.

For more details and interpretation, see the vignette by: browseVignettes("ReQON").

Value

In addition to the plot, FWSEplot calculates and outputs Frequency-Weighted Squared Error (FWSE), a measure of how close the points lie to the 45-degree line.

Author(s)

Christopher Cabanski <cabanski@email.unc.edu>

Examples

```
## Create relative frequency example data
require( stats )
after <- dnorm( c( 0:40 ), mean = 30, sd = 8 )
  after <- after / sum( after )
err_rate <- c( 0:40 ) + rnorm( 41, mean = 0, sd = 5)
  err_rate[ which( err_rate < 0 ) ] <- (-1) * err_rate[ which( err_rate < 0 ) ]
  ## to guarantee that all values are positive

## plot and calculate FWSE
FWSEplot( err_rate, after, col = "red" )
```

QualFreqPlot *Plot frequency distributions of quality scores.*

Description

Plots the relative frequency distribution of quality scores before and after recalibration.

Usage

```
QualFreqPlot(QualFreqBefore, QualFreqAfter, before_col = "blue",
   after_col = "red", inc_legend = TRUE, xlabel = "Quality Score",
   ylabel = "Relative Frequency",
   main_title = "Frequency Distributions of Quality Scores")
```

Arguments

QualFreqBefore	vector of relative frequencies of quality scores before recalibration. The first element in the vector corresponds to a quality score of zero.
QualFreqAfter	vector of relative frequencies of quality scores after recalibration. The first element in the vector corresponds to a quality score of zero.
before_col	color of line plotting the frequency before recalibration. Default = "blue".
after_col	color of line plotting the frequency after recalibration. Default = "red".
inc_legend	option for including a legend. Default = TRUE.
xlabel	x-axis label. Default = "Quality Score".
ylabel	y-axis label. Default = "Relative Frequency".
main_title	title. Default = "Frequency Distributions of Quality Scores".

Details

QualFreqPlot plots the relative frequency distribution of quality scores before and after recalibration. If the input vectors are $QualFreqBefore and $QualFreqAfter from ReQON output, this function will create the top right diagnostic plot that is output from ReQON.

For more details and interpretation, see the vignette by: browseVignettes("ReQON").

Author(s)

Christopher Cabanski <cabanski@email.unc.edu>

Examples

```
## Create data of frequencies
require( stats )
before <- dpois( c( 0:40 ), 40 )
  before <- before / sum( before )
after <- dnorm( c( 0:40 ), mean = 30, sd = 8 )
  after <- after / sum( after )

## plot
QualFreqPlot( before, after )
```

ReQON *Recalibrating Quality Of Nucleotides*

Description

Recalibrate the nucleotide quality scores of either single-end or paired-end next-generation sequencing data that has been aligned.

Usage

```
ReQON(in_bam, out_bam, region, max_train = -1, SNP = "",
   RefSeq = "", plotname = "", temp_files = 0)
```

Arguments

in_bam	file name of sorted BAM file of single-end or paired-end aligned sequencing data. The corresponding index file (.bai file) must be located in the same directory.
out_bam	file name for output BAM file with original quality scores replaced with recalibrated quality scores.
region	training region for recalibration, as "chromosome:start-end". Cannot span more than one chromosome. See note. Example: "chr1:1-10000".
max_train	maximum number of nucleotides to include in training region. Useful if you want to train on e.g. the first 5 million bases of chromosome 10. Default = -1 (use all nucleotides from training region).

`SNP` file of SNP locations to remove from training set before recalibration. Text or Rdata file (with variable name "snp") with no header and two columns: [1] chromosome, [2] position. See note.

Default: do not remove any nucleotides from training set.

`RefSeq` file of reference sequence for training set to identify sequencing errors (i.e, nucleotide is error if it does not match RefSeq). Text or Rdata file (with variable name "ref") with no header and three columns: [1] chromosome, [2] position, [3] reference nucleotide (A,C,G,T). See note.

Default: errors are nucleotides not matching major allele(s) for coverage > 2, removing all nucleotides at positions with coverage of 2 or less.

`plotname` file name for saving recalibration plots in pdf. If not specified, plots will not be produced.

`temp_files` option for keeping temporary files.

0: (default) remove all temporary files.

1: keep temporary files in working directory.

Details

ReQON uses logistic regression to recalibrate the nucleotide quality scores of a sorted BAM file. The BAM file contains either single-end or paired-end next-generation sequencing data that has been aligned using any alignment tool. For help with sorting and indexing BAM files in R, see Rsamtools.

ReQON also has the option to output diagnostic plots which show the effectiveness of the recalibration on the training set.

For a detailed description of usage, output and images, see the vignette by: browseVignettes("ReQON").

ReQON utilizes various java tools provided by Picard. For more information on Picard, see http://picard.sourceforge.net

Value

ReQON returns a BAM file, replacing the original quality scores with the recalibrated quality scores in the QUAL field.

ReQON also outputs a data object of diagnostic data from the training set that is plotted in the output diagnostic plots. The object variables are:

`$ReadPosErrors`
: vector of error counts by read position.

`$QualFreqBefore`
: relative frequency of quality scores before recalibration. The first element in the vector corresponds to a quality score of zero.

`$QualFreqAfter`
: relative frequency of quality scores after recalibration. The first element in the vector corresponds to a quality score of zero.

`$ErrRatesBefore`
: vector of empirical error rates before recalibration, reported on the Phred scale. The first element in the vector corresponds to a quality score of zero.

`$ErrRatesAfter`
: vector of empirical error rates after recalibration, reported on the Phred scale. The first element in the vector corresponds to a quality score of zero.

`FWSE` vector of Frequency-Weighted Squared Error (FWSE) values. The first element is FWSE before recalibration and the second element is FWSE after recalibration.

Note

Be aware of how the chromosomes are referenced when specifying the training region. For example, one BAM file may require specifying "10:1-2000" while another may need "chr10:1-2000".

If providing SNP or RefSeq files, computations will speed up if your file only covers the positions in the training region. For example, if you set region = "chr10:1-2000", then we recommend only having rows corresponding to chr10:1-2000 in the RefSeq/SNP file.

Author(s)

Christopher Cabanski <cabanski@email.unc.edu>

Examples

```
## Read in sample data from seqbias package
library( ReQON )
library( seqbias )
library( Rsamtools )
ref_fn <- system.file( "extra/example.fa", package = "seqbias" )
ref_f <- FaFile( ref_fn )
open.FaFile( ref_f )
reads_fn <- system.file( "extra/example.bam", package = "seqbias" )

## Set up file of reference sequence
seqs <- scanFa( ref_f )
len <- length( seqs[[1]] )
ref <- matrix( nrow = len, ncol = 3 )
ref[,1] <- rep( "seq1", len )
ref[,2] <- c( 1:len )
str <- toString( subseq( seqs[[1]], 1, len ) )
s <- strsplit( str, NULL )
ref[,3] <- s[[1]]
write.table( ref, file = "ref_seq.txt", sep = "\t", quote = FALSE,
   row.names = FALSE, col.names = FALSE )

## Recalibrate File
sorted <- sortBam( reads_fn, tempfile() )
indexBam( sorted )
reg <- paste( "seq1:1-", len, sep = "" )
diagnostics <- ReQON( sorted, "Recalibrated_example.bam", reg,
   RefSeq = "ref_seq.txt", plotname = "Recalibrated_example_plots.pdf" )

#Remove temporary file
unlink( "ref_seq.txt" )
```

ReadPosErrorPlot *Plot distribution of errors by read position.*

Description

Plots the number of sequencing errors by read position.

Usage

```
ReadPosErrorPlot(ReadPosErrors, error_col = "blue", thresh = 1.5,
   thresh_col = "cyan", xlabel = "Read Position", ylabel = "# Errors",
   main_title = "Distribution of Errors by Read Position")
```

Arguments

ReadPosErrors	vector of sequencing error counts by read position.
error_col	color of line plotting the errors counts. Default = "blue".
thresh	Threshold for identifying read positions with large numbers of errors, plotted as a horizontal dashed line. Threshold is set as "thresh * (average number of errors per read position)". Default = 1.5.
thresh_col	color of threshold line. Default = "cyan".
xlabel	x-axis label. Default = "Read Position".
ylabel	y-axis label. Default = "# Errors".
main_title	title. Default = "Distribution of Errors by Read Position".

Details

ReadPosErrorPlot plots the distribution of sequencing errors by read position. If the input vector is $ReadPosErrors from ReQON output, this function will create the top left diagnostic plot that is output from ReQON.

For more details and interpretation, see the vignette by: browseVignettes("ReQON").

Author(s)

Christopher Cabanski <cabanski@email.unc.edu>

Examples

```
## Create data of error counts
x <- c( 1:30 )
err <- x^2 + ( 30 - x )^1.6 + rnorm(30, 0, 100)

## plot errors by read position
ReadPosErrorPlot( err )
```

Index