Package ‘Rariant’

March 23, 2017

Type Package

Title Identification and Assessment of Single Nucleotide Variants through Shifts in Non-Consensus Base Call Frequencies

Version 1.10.0

Author Julian Gehring, Simon Anders, Bernd Klaus

Maintainer Julian Gehring <jg-bioc@gmx.com>

Imports methods, S4Vectors, IRanges, GenomeInfoDb, ggbio, ggplot2, exomeCopy, SomaticSignatures, Rsamtools, shiny, VGAM, dplyr, reshape2

Depends R (>= 3.0.2), GenomicRanges, VariantAnnotation

Suggests h5vcData, testthat, knitr, optparse, BSgenome.Hsapiens.UCSC.hg19

Description The ‘Rariant’ package identifies single nucleotide variants from sequencing data based on the difference of binomially distributed mismatch rates between matched samples.

VignetteBuilder knitr

Encoding UTF-8

ByteCompile TRUE

License GPL-3

URL https://github.com/juliangehring/Rariant

BugReports https://support.bioconductor.org

LazyLoad yes

biocViews Sequencing, StatisticalMethod, GenomicVariation, SomaticMutation, VariantDetection, Visualization

NeedsCompilation no

R topics documented:

Rariant-package ................................................................. 2
  ciAdjust ............................................................... 3
  ciAssessment .......................................................... 3
  ciUtils ................................................................. 4
  colorscales ............................................................. 5
The 'Rariant' package offers the framework to identify and characterize shifts of variant frequencies in a comparative setting from high-throughput short-read sequencing data. It estimates shifts in the non-consensus variant frequency and provides confidence estimates that allow for a quantitative assessment of presence or absence of variants. The vignette accompanying the package gives a detailed explanation and outlines a typical workflow on real data.

Author(s)

Julian Gehring, Simon Anders, Bernd Klaus (EMBL Heidelberg)

Maintainer: Julian Gehring <julian.gehring@embl.de>

See Also

rariant

vignette(package = "Rariant")

Examples

help("rariant")

vignette(package = "Rariant")
ciAdjust

Description

Multiple testing adjustment of confidence levels, as proposed by Benjamini and Yekutieli.

Usage

\[
\text{ciAdjustLevel}(\eta_0, \text{conf}\_\text{level})
\]

Arguments

- \(\eta_0\) Estimated fraction of tests that are consistent with the null hypothesis.
- \(\text{conf}\_\text{level}\) Unadjusted confidence level

Value

The adjusted confidence level.

References


Examples

\[
\begin{align*}
\text{conf}\_\text{level} &= 0.95 \\
\eta_0 &= \text{seq}(0, 1, \text{by} = 0.02) \\
\text{conf}\_\text{level}\_\text{adj} &= \text{ciAdjustLevel}(\eta_0, \text{conf}\_\text{level}) \\
\text{plot}(\eta_0, \text{conf}\_\text{level}\_\text{adj}, \text{pch} = 20, \text{ylim} = \text{c(conf}\_\text{level}, 1))
\end{align*}
\]

ciAssessment

Description

Functions to compute the coverage probability of a confidence interval method.

Usage

\[
\text{coverageProbability(pars, fun = acCi, n}\_\text{sample} = 1e4, \text{min}\_k, \ldots)
\]
Arguments

pars Data frame with parameter combinations [data.frame]
n_sample Number of assessments per parameter combination [integer(1)].
fun CI function
min_k Minimum 'k2' value to use.
... Additional arguments that are passed on to 'fun'.

Value

The 'data.frame' object 'pars' with additional columns 'cp' for the coverage probability and 'aw' average confidence interval width.

References


Examples

## Define parameter space
pars = expand.grid(k1 = 1:5, k2 = 5, n1 = 30, n2 = 30)
conf_level = 0.95
## Compute coverage probabilities
cp = coverageProbability(pars, fun = acCi, n_sample = 1e2, conf_level = conf_level)
print(cp)

Description

Utility functions to find confidence intervals that (a) overlap a certain value ('ciOutside', 'ciCovers') and (b) different confidence intervals overlap ('ciOverlap').

Usage

    ciOutside(x, delta = 0)
    ciCovers(x, delta = 0)
    ciOverlap(x, y)
    ciWidth(x)

Arguments

x, y CIs, as obtained from e.g. the 'acCi' function.
delta Variant frequency value to check against [default: 0].
colorscales

Value
A logical vector, where each elements corresponds to the respective row of 'x' (and 'y').
For 'ciWidth': A numeric vector with the widths of the confidence intervals.

Examples

```r
## Generate sample data
counts = data.frame(x1 = 1:5, n1 = 30, x2 = 0:4, n2 = 30)

## Agresti-Caffo
ci_ac = with(counts, acCi(x1, n1, x2, n2))
ci_ac2 = with(counts, acCi(x1, n1, x2, n2, 0.99))

## cover 0
idx_zero = ciCovers(ci_ac)

## cover 1
idx_one = ciCovers(ci_ac, delta = 1)

## overlap
idx_same = ciOverlap(ci_ac, ci_ac2)

## width
width = ciWidth(ci_ac)
```

---

colorscales  |  Rariant color scales

Description
Color and fill scales used for plotting in Rariant.

Usage

```r
eventFillScale()
verdictColorScale()
rateFillScale()
baseFillScale()
```

Value
A ggbio color or fill scale.

See Also
tallyPlot, evidenceHeatmap, plotConfidenceIntervals
convertUtils  Position converters

Description
Utility functions to convert between 'GRanges' and 'character' objects.

Usage

```r
gr2pos(x, range = TRUE)
pos2gr(x)
```

Arguments

- `x`: GRanges or character object.
- `range`: Should the range instead of the start position be returned?

Value
A GRanges object or character object, with the positions.

Examples

```r
library(GenomicRanges)
gr = GRanges(1:2, IRanges(1:2, width = 1))
pos = gr2pos(gr)
gr2 = pos2gr(pos)
identical(gr, gr2)
```

evidenceHeatmap  Variant Evidence Heatmap

Description
Heatmap with the evidence of variant evidences.

Usage

```r
evidenceHeatmap(x, fill = "d", color = "outside",
                 size = 1.5,
                 xvar = "sample", yvar = "loc", ...)
```
mismatchUtils

Arguments

x
fill
color
size
xvar, yvar
...

GRanges with variants, as returned by 'rvariant'.
Column determining the fill.
Column determining the border color.
What is this needed for?
Which column to define the tiles.
Additional arguments, passed to 'geom_tile'.

Value

A ggplot2 object.

See Also

yesNoMaybe

mismatchUtils  Tally processing low-level functions

Description

Functions for processing position-specific base count tables (tallies) and extracting mismatches counts.

Usage

## low-level functions
selectStrand(x, strand = c("both", "plus", "minus"), idx = 1:ncol(x))
seqDepth(x)
callConsensus(counts, verbose = FALSE)
mismatchCount(counts, consensus, depth = rowSums(counts))

Arguments

x
strand
idx
counts
verbose
consensus
depth

Input object
Which strand to return?
Index of bases to consider (leave as is)
Count matrix
Show warnings
Consensus sequence
Sequencing depth for counts.

See Also

comparativeMismatch
multiCalls

Multi call processing

Description
Utilities for processing matched calls from multiple samples.

Usage

findCalls(x, ..., minCount = 1)
filterCalls(x, ..., minCount = 1)
mergeCalls(x)
updateCalls(x, ...)

Arguments
x  GenomicRangesList with calls from multiple samples.
... Additional arguments.
minCount For finding and filtering, for how many samples must the condition ‘...’ hold true for a site to be returned?

plotCIs

Plotting Functions

Description
The ‘plotConfidenceIntervals’ is a high-level plotting function for visualizing confidence intervals. The ‘plotAbundanceShift’ function visualizes the shift in mismatch rates between two samples.

Usage

plotConfidenceIntervals(x, ylim = c(-1.05, 1.05), color = NULL, ...)

plotAbundanceShift(x, ylim = c(-0.05, 1.05), rates = TRUE, ...)

Arguments
x  ‘GRanges’ with mcols of a CI method, or ‘data.frame’ as returned by one of the CI methods, with the optional column ‘start’.
ylim Limits of the y-axis. Using this instead of using the ‘ylim’ prevents ugly warnings of ‘ggplot2’.
color Variable that determines the coloring of the confidence axis (character).
rates Should the non-consensus rates of both samples be visualized as colored end points of the line range? (logical, default: TRUE).
... Additional plotting arguments that are passed on to ggplot2::geom_pointrange.
propCIs

Value

For a 'GRanges' input: A ‘ggbio’ object
For a 'data.frame' input: A ‘ggplot’ object

Examples

```r
## Generate sample data
counts = data.frame(x1 = 1:5, n1 = 30, x2 = 0:4, n2 = 30)

## Agresti-Caffo
ci_ac = with(counts, acCi(x1, n1, x2, n2))

library(GenomicRanges)
gr = GRanges("1", IRanges(start = 1:nrow(counts), width = 1))
mcols(gr) = ci_ac

## GRanges
plotConfidenceIntervals(gr)

## data.frame
plotConfidenceIntervals(ci_ac)

## abundance shift
plotAbundanceShift(gr)
plotAbundanceShift(ci_ac)
```

---

**propCIs**  
*Confidence Interval Functions*

Description

Vectorized implementation of confidence intervals

Usage

```r
acCi(x1, n1, x2, n2, conf_level = 0.95, clip = TRUE, split = FALSE)
nhsCi(x1, n1, x2, n2, conf_level = 0.95)
```

Arguments

- `x1`: Mismatch counts in the test sample.
- `n1`: Sequencing depth (total counts) in the test sample.
- `x2`: Mismatch counts in the control sample.
- `n2`: Sequencing depth (total counts) in the control sample.
- `conf_level`: Confidence level $\beta$ (default: 0.95).
- `clip`: Should the CIs be clipped to the interval [-1,1] if they exceed this?
- `split`: Should the sample split method be applied? See ‘splitSampleBinom’ for details.
**Details**

These functions implement a vectorized version of the two-sided Agresti-Caffo, and Newcombe-Hybrid-Score confidence interval for the difference of two binomial proportions.

**Value**

A data frame with columns

- dEstimate for the difference of rates 'p1' and 'p2'.
- p1, p2Estimates for the mismatches rates for each sample.
- lower, upperLower and upper bound of the confidence interval.
- wWidth of the confidence interval.

**References**


**See Also**

nhsCi

splitSampleBinom

binMto::Add4 binMto::NHS

**Examples**

```r
## Generate sample data
counts = data.frame(x1 = 1:5, n1 = 30, x2 = 0:4, n2 = 30)

## Agresti-Caffo
ci_ac = with(counts, acCi(x1, n1, x2, n2))

## Newcombe-Hybrid Score
ci_nhs = with(counts, nhsCi(x1, n1, x2, n2))

print(ci_ac)
```
propTests

| propTests | Testing Functions |

Description
Vectorized implementation of testing functions

Usage

\[\text{scoreTest}(x_1, n_1, x_2, n_2)\]
\[\text{nmTest}(x_1, n_1, x_2, n_2, \text{delta} = 0)\]
\[\text{feTest}(x_1, n_1, x_2, n_2, \ldots)\]

Arguments

\[x_1\] Mismatch counts in the test sample.
\[n_1\] Sequencing depth (total counts) in the test sample.
\[x_2\] Mismatch counts in the control sample.
\[n_2\] Sequencing depth (total counts) in the control sample.
\[\text{delta}\] Difference to test against (default: 0).
\[\ldots\] Additional arguments.

Details
These functions implement a vectorized version of the two-sided (a) Score test and (b) Miettinen-Nurminen test for the difference between two Binomial proportions. Usage of the score test is discouraged in the settings considered here, since it is ill-defined for positions with no mismatches.

Value
A data frame with columns

- \(\text{dhatEstimate}\) for the difference of rates \('p1' and 'p2'.
- \(p1, p2\)Estimates for the mismatches rates for each sample.
- \(t\text{val}\)T-value
- \(p\text{val}\)P-value

References


See Also

VariantTools package
Examples

```r
## Generate sample data
counts = data.frame(x1 = 1:5, n1 = 30, x2 = 0:4, n2 = 30)

## Score test
stat_st = with(counts, scoreTest(x1, n1, x2, n2))

## NM test
stat_nm = with(counts, nmTest(x1, n1, x2, n2))

## Fisher test
stat_fet = with(counts, feTest(x1, n1, x2, n2))

print(stat_st)
print(stat_nm)
print(stat_fet)
```

rariant

rariant calling functions

Description

The ‘rariant’ function identifies variant shifts between a test and control sample. These high-level functions offer a convenient interface for large-scale identification as well as for reexamination of existing variant calls.

Usage

```r
## S4 method for signature 'BamFile,BamFile,GRanges'
rariant(test, control, region,
beta = 0.95, alpha = 1 - beta, select = TRUE, consensus,
resultFile, strand = c("both", "plus", "minus"), nCycles = 10,
minQual = 20, block = 1e4, value = TRUE, criteria = c("both",
"any", "fet", "ci"))

## S4 method for signature 'character,character,GRanges'
rariant(test, control, region,
beta = 0.95, alpha = 1 - beta, select = TRUE, consensus,
resultFile, strand = c("both", "plus", "minus"), nCycles = 10,
minQual = 20, block = 1e4, value = TRUE, criteria = c("both",
"any", "fet", "ci"))

## S4 method for signature 'array,array,GRanges'
rariant(test, control, region, beta = 0.95, alpha = 1 - beta, select = TRUE, consensus, strand = c("both", "plus", "minus"), criteria = c("both", "any", "fet", "ci"))

rariantStandalone()
```
readRariant(file, ...)

writeRariant(x, file, ...)

Arguments

test, control  Test and control BAM files. Other input sources will be supported in the future.
region  Region(s) of interest to analyze in the calling [GRanges with one or more entries]. If missing, the entire genomic space, as defined by the BAM headers of the input files, will be covered.
beta  Confidence level [numeric in the range [0,1], default: 0.95].
alpha  Significance threshold for BH-adjusted p-values of the Fisher’s exact test.
select  Should only likely variant positions be selected and returned, or the results for all sites be returned.
consensus  How to determine the consensus sequence. By default, the consensus is given by the most abundant allele in the control sample. Alternatively, an object with a reference sequence (‘BSgenome’, ‘FaFile’) can be passed to define the consensus sequence.
resultFile  If not missing, write the results to a tab-delimited file.
strand  Which strand should be extracted? By default, the counts of both strands are summed up.
nCycles  Number of sequencing cycles to remove from the beginning and end of each read when creating the base count table. This avoids low quality read positions [default: 10 is reasonable for current Illumina sequencing].
minQual  Minimum base call quality for reads to be considered for the nucleotide count table [default: 20 is reasonable for current Illumina sequencing]. Reads with a lower quality are dropped.
block  Number of the genomic sites to analyze in one chunk. The default is a good compromise between memory usage and speed, and normally does not require changing.
value  Should the results be returned by the function. For calls within R, this is generally set to TRUE and does not need to be changed.
criteria  The criteria to determine significant sites. Criteria are: Fisher’s exact test, confidence intervals, any or both [default] of them.

Details

The ‘rariant’ function is the workhorse for the comparative variant calling and assessment. It starts with the aligned reads for the test (e.g. tumor) and the control (e.g. normal) sample in the BAM format; later versions will support additional inputs.

The ‘select’ parameter determines whether only significant variant sites or all sites are returned. While the first is suitable for detecting variants, the second becomes relevant assessing for example the abundance of variants at particular sites of interest - an example would be to determine the absence of a specific variant.
For analyses over large genomic regions and for use with infrastructure outside of R, initiating the calling from the command line may be a desirable alternative. The ‘rariantStandalone’ functions returns the full path to a script that can be directly called from the command line. For further details, see the help of the script by calling it with the ‘-h’ option, for example ‘rariant -h’.

The ‘readRariant’ and ‘writeRariant’ functions allow to import and export the results of a ‘rariant’ call from and to a file output, and will return the same object.

Value

A `GRanges` object, with each row corresponding to a genomic site, and columns:

- testMismatch, controlMismatchMismatch counts in the test and control sample.
- testDepth, controlDepthSequencing depth in the test and control sample.
- testRef, testAltReference and alternative allele of the test sample.
- controlRefReference allele of the control sample.
- testRefDepth, testAltDepthSupporting sequencing depth for the reference and alternative allele in the test sample.
- refConsensus allele.
- p1, p2Estimated non-consensus rate in test and control, respectively.
- dEstimated shift in the non-consensus rate between test and control.
- dsEstimated shift in the non-consensus rate between test and control (shrinkage point estimate).
- lower, upperLower and upper bound of the confidence interval for ‘d’.
- pval, padjRaw and BH-adjusted p-value of the FET test.
- calledWas the site identified as variant?
- eventTypeThe class of the event: somatic, heterozygous, undecided.
- padjSomatic, padjHeteroBH-adjusted p-values of the binomial tests for ‘eventType’.
- pvalSomatic, pvalHeteroRaw p-values of the binomial tests for ‘eventType’.

Examples

```r
library(GenomicRanges)

control_bam = system.file("extdata", "NRAS.Control.bam", package = "h5vcData", mustWork = TRUE)
test_bam = system.file("extdata", "NRAS.AML.bam", package = "h5vcData", mustWork = TRUE)

roi = GRanges("1", IRanges(start = 115258439, end = 115259089))

vars = rariant(test_bam, control_bam, roi)
vars_all = rariant(test_bam, control_bam, roi, select = FALSE)

## Not run:
system2(rariantStandalone(), "-h")

## End(Not run)
```
**rariantInspect**  
*Interactive inspection*

**Description**
Interactively inspect variant sites and results of the `rariant` function.

**Usage**

```r
rariantInspect(x)
```

**Arguments**

`x`  
The return value of the `rariant` or `readRariant` function.

**Details**
With the web interface of `rariantInspect` can existing variant calls and assessment be explored interactively. It allows to select the genomic region of interest and the type of event. Results are shown as both a confidence interval plot and a results table that can be further filtered and reordered.

**Examples**

```r
example(rariant)
rariantInspect(vars_all)
```

---

**splitSample**  
*Split Sample for Binomial Data*

**Description**
Sample splitting, according to Hall, 2014.

**Usage**

```r
splitSampleBinom(x, n)
```

**Arguments**

`x`  
Number of successes

`n`  
Number of trials
Details

These functions implement sample splitting of a binomial rate.

Note that the results depend on the state of the random number generator, and are therefore not strictly deterministic.

Value

A vector with the rate \( p = \frac{X}{N} \), obtained with sample splitting.

References


Examples

```r
n = 10
m = 5
pt = 0.5
x = rbinom(m, n, pt)
p = x/n
ps = splitSampleBinom(x, n)
round(cbind(p, ps), 2)
```

---

tallyBam

*Tally a genomic region*

Description

Create the nucleotide count table (‘tally’) of a genomic region from a BAM file.

Usage

`tallyBamRegion(bam, region, minBase = 0, minMap = 0, maxDepth = 10000)`

Arguments

- `bam` BAM file
- `region` GRanges with the region to tally, with one entry.
- `minBase`, `minMap` Minimum base call and mapping quality for reads to be considered for the nucleotide count table [default: 0]. Reads with a lower quality are dropped.
- `maxDepth` Maximal sequencing depth to analyze.

Details

For details, look at the documentation of the underlying 'tallyBAM' function in the 'h5vc' package.
tallyPlot

Value

An integer array with the dimensions:

- positionLength: width(region)
- baseA, C, G, T
- strand+, -

See Also

h5vc::tallyBAM, deepSNV::bam2R, Rsamtools::pileup

tallyPlot  
Mismatch plot from BAM files

Description

Create a mismatch plot from a list of BAM files directly.

Usage

tallyPlot(file, region, ref, nCycles = 0, minQual = 0, minFreq = 0, ...)

Arguments

- file: BAM file paths
- region: GRanges with the position (width: 1) to tally, with one entry.
- ref: Reference object, as ‘BSgenome’.
- nCycles: Number of sequencing cycles to remove from the beginning and end of each read when creating the base count table. This avoids low quality read positions [default: 0]. See ‘tallyBamRegion’
- minQual: Minimum base call quality for reads to be considered for the nucleotide count table [default: 0]. Reads with a lower quality are dropped. See ‘tallyBamRegion’
- minFreq: Currently not used
- ...: Additional arguments, passed to ‘tallyBAM’.

Value

A ‘ggplot2’ or ‘ggbio’ object.

See Also

h5vc::mismatchPlot
Examples

```r
library(ggbio)
library(GenomicRanges)
library(BSgenome.Hsapiens.UCSC.hg19)

region = GRanges(\"chr17\", IRanges(7572100, width = 1))

control_bam = system.file("extdata", "platinum", "control.bam", package = "Rariant", mustWork = TRUE)
mix_bam = system.file("extdata", "platinum", "mix.bam", package = "Rariant", mustWork = TRUE)

bam_files = c(control_bam, mix_bam)

region = GRanges(\"chr17\", IRanges(7572050, width = 100))

correct_bam = system.file("extdata", "Correct", "control.bam", package = "Rariant", mustWork = TRUE)
test1_bam = system.file("extdata", "platinum", "test.bam", package = "Rariant", mustWork = TRUE)
test2_bam = system.file("extdata", "platinum", "test2.bam", package = "Rariant", mustWork = TRUE)
mix_bam = system.file("extdata", "platinum", "mix.bam", package = "Rariant", mustWork = TRUE)

correct_files = c(correct_bam, test1_bam, test2_bam, mix_bam)

library(BSgenome.Hsapiens.UCSC.hg19)
ref = BSgenome.Hsapiens.UCSC.hg19

p = tracks(lapply(bam_files, tallyPlot, region, ref, minQual = 25))

print(p)
```

---

**yesNoMaybe**

**Determine Variant Evidence**

**Description**

Determine the evidence (absence, presence, don't know) of variants.

**Usage**

```
yesNoMaybe(x, null = 0, one = 0.5)
```

**Arguments**

- **x**: GRanges with variants, as returned by `rariant`.
- **null**: Shift consistent with the _absence_ of a variant.
- **one**: Shift consistent with the _presence_ of a variant.
Value

The same GRanges object as the input 'x', with the factor column 'verdict': 'absent', 'present', 'inbetween', 'dontknow'
Index

*Topic package
  Rariant-package, 2
acCi (propCIs), 9
baseFillScale (colorscales), 5
callConsensus (mismatchUtils), 7
ciAdjust, 3
ciAdjustLevel (ciAdjust), 3
ciAssessment, 3
ciCovers (ciUtils), 4
ciOutside (ciUtils), 4
ciOverlap (ciUtils), 4
ciUtils, 4
ciWidth (ciUtils), 4
colorscales, 5
convertUtils, 6
coverageProbability (ciAssessment), 3
eventFillScale (colorscales), 5
evidenceHeatmap, 6
feTest (propTests), 11
filterCalls (multiCalls), 8
findCalls (multiCalls), 8
gr2pos (convertUtils), 6
mergeCalls (multiCalls), 8
mismatchCount (mismatchUtils), 7
mismatchUtils, 7
multiCalls, 8
nhsCi (propCIs), 9
nmTest (propTests), 11
plotAbundanceShift (plotCIs), 8
plotCIs, 8
plotConfidenceIntervals (plotCIs), 8
pos2gr (convertUtils), 6
propCIs, 9
propTests, 11
Rariant (Rariant-package), 2
rariant, 12
rariant, array, array, GRanges-method (rariant), 12
rariant, BamFile, BamFile, GRanges-method (rariant), 12
rariant, character, character, GRanges-method (rariant), 12
rariant-methods (rariant), 12
Rariant-package, 2
rariantInspect, 15
rariantStandalone (rariant), 12
rateFillScale (colorscales), 5
readRariant (rariant), 12
scoreTest (propTests), 11
selectStrand (mismatchUtils), 7
seqDepth (mismatchUtils), 7
splitSample, 15
splitSampleBinom (splitSample), 15
tallyBam, 16
tallyBamRegion (tallyBam), 16
tallyPlot, 17
updateCalls (multiCalls), 8
verdictColorScale (colorscales), 5
writeRariant (rariant), 12
yesNoMaybe, 18