Package 'SimFFPE'

October 17, 2020

Type Package

Title NGS Read Simulator for FFPE Tissue

Version 1.0.0

Author Lanying Wei

Maintainer Lanying Wei <lanying.wei@uni-muenster.de>

Description This package simulates artifact chimeric reads specifically generated in next-generation sequencing (NGS) process of formalin-fixed paraffin-embedded (FFPE) tissue.

License LGPL-3

Encoding UTF-8

Depends Biostrings

Imports dplyr, foreach, doParallel, truncnorm, GenomicRanges, IRanges, Rsamtools, parallel, graphics, stats, utils, methods

Suggests BiocStyle

biocViews Sequencing, Alignment, MultipleComparison, SequenceMatching, DataImport

git_url https://git.bioconductor.org/packages/SimFFPE

git_branch RELEASE_3_11

git_last_commit 892fed2

git_last_commit_date 2020-04-27

Date/Publication 2020-10-16

R topics documented:

SimFFPE-package *NGS Read Simulator for FFPE Tissue*

Description

This package simulates artifact chimeric reads specifically generated in next-generation sequencing (NGS) process of formalin-fixed paraffin-embedded (FFPE) tissue.

Details

This package was not yet installed at build time.
The NGS (Next-Generation Sequencing) reads from FFPE (Formalin-Fixed Paraffin-Embedded) samples contain numerous artificial chimeric reads. These reads are derived from the combination of two single-stranded DNA (ss-DNA) fragments with short reverse complementary sequences. The combined ss-DNA may come from adjacent or distant regions. This package simulates these artifacts as well as normal reads for FFPE samples. The simulation can cover whole genome, or several chromosomes, or large regions, or whole exome, or targeted regions. It also supports enzymatic / random fragmentation and paired-end / single-end sequencing simulations. Fine-tuning can be performed for desired simulation results, and multi-threading can help reduce the runtime. Please check the package vignette for the guidance of fine-tuning.

Index: This package was not yet installed at build time.
There are three available functions for NGS read simulation of FFPE samples:

1. calcPhredScoreProfile: Calculate positional Phred score profile from BAM file for read simulation.

2. readSimFFPE: Simulate noisy NGS reads of FFPE samples on whole genome, or several chromosomes, or large regions.

3. targetReadSimFFPE: Simulate noisy NGS reads of FFPE samples in exonic / targeted regions.

Author(s)

Lanying Wei

Maintainer: Lanying Wei <lanying.wei@uni-muenster.de>

See Also

calcPhredScoreProfile, readSimFFPE, targetReadSimFFPE

Examples

```
PhredScoreProfilePath <- system.file("extdata", "PhredScoreProfile2.txt",
                                     package = "SimFFPE")
PhredScoreProfile <- as.matrix(read.table(PhredScoreProfilePath, skip = 1))
colnames(PhredScoreProfile) <- read.table(PhredScoreProfilePath,
                                          nrows = 1,
                                          colClasses = "character")

referencePath <- system.file("extdata", "example.fasta", package = "SimFFPE")
reference <- readDNAStringSet(referencePath)

## Simulate reads of the first three sequences of the reference genome
```

```
sourceSeq <- reference[1:3]
outFile1 <- paste0(tempdir(), "/sim1")
readSimFFPE(sourceSeq, referencePath, PhredScoreProfile, outFile1,
            coverage = 80, enzymeCut = TRUE, threads = 4)

## Simulate reads for targeted regions

bamFilePath <- system.file("extdata", "example.bam", package = "SimFFPE")
regionPath <- system.file("extdata", "regionsBam.txt", package = "SimFFPE")
regions <- read.table(regionPath)
PhredScoreProfile <- calcPhredScoreProfile(bamFilePath, targetRegions = regions)

regionPath <- system.file("extdata", "regionsSim.txt", package = "SimFFPE")
targetRegions <- read.table(regionPath)

outFile <- paste0(tempdir(), "/sim2")
targetReadSimFFPE(referencePath, PhredScoreProfile, targetRegions, outFile,
                  coverage = 120, readLen = 100, meanInsertLen = 150,
                  sdInsertLen = 40, enzymeCut = FALSE)
```

`calcPhredScoreProfile` *Estimate Phred score profile for FFPE read simulation*

Description

Calculate Phred score profile from the entire BAM file or reads in subsampled regions.

Usage

```
calcPhredScoreProfile(bamFilePath, mapqFilter = 0, maxFileSize = 1,
targetRegions = NULL, subsampleRatio = NA, subsampleRegionLength = 1e+05,
disableSubsampling = FALSE, threads = 1)
```

Arguments

`bamFilePath`	BAM file to be processed.
`mapqFilter`	Filter for mapping quality. Reads with mapping quality below this value will be excluded from calculation.
`maxFileSize`	The maximum file size (in GB) that allows processing of the entire BAM file. If disableSubsampling is set to false, BAM file larger than this size will be subsampled for calculation.
`targetRegions`	A DataFrame or GenomicRanges object representing target regions for calculation. Use it for targeted sequencing / WES data, or when you need to manually select subsampled regions (set disableSubsampling to true in this case). If it is a DataFrame, the first column should be the chromosome, the second the start position and the third the end position. Please use one-based coordinate systems (the first base should be marked with 1 but not 0).
`subsampleRatio`	Subsample ratio. Together with subsampleRegionLength to determine subsampled regions. When subsampleRatio is not given, it will be assigned the value of maxFileSize divided by the input BAM file size. Range: 0 to 1.

`subsampleRegionLength`
: Length of each subsampled region. Unit: base pair (bp).

`disableSubsampling`
: Force to use the entire BAM file for calculation when set to true.

`threads`
: Number of threads used. Multi-threading can speed up the process.

Details

Calculate positional Phred score profile from the entire BAM file or reads in subsampled regions. A Phred score profile will be returned, which can then be used in read simulation.

Value

A matrix will be returned. Each row of the matrix represents a position in the read (from begin to end), and each column the Phred quality score of base-calling error probabilities. The value in the matrix represents the positional Phred score proportion.

Author(s)

Lanying Wei <lanying.wei@uni-muenster.de>

See Also

`SimFFPE`, `readSimFFPE`, `targetReadSimFFPE`

Examples

```
bamFilePath <- system.file("extdata", "example.bam", package = "SimFFPE")
regionPath <- system.file("extdata", "regionsBam.txt", package = "SimFFPE")
regions <- read.table(regionPath)
PhredScoreProfile <- calcPhredScoreProfile(bamFilePath, targetRegions = regions)
```

---

readSimFFPE *Simulate noisy NGS reads of FFPE samples for whole genome / several chromosomes / large regions*

Description

NGS data from FFPE samples contain numerous artificial chimeric reads. These chimeric reads are formed through the combination of two single-stranded DNA (ss-DNA). This function simulates these artificial reads as well as normal reads for FFPE samples on whole genome, or several chromosomes, or large regions.

Usage

```
readSimFFPE(sourceSeq, referencePath, PhredScoreProfile, outFile, coverage,
readLen = 150, meanInsertLen = 250, sdInsertLen = 80, enzymeCut = FALSE,
chimericRatio = 0.08, localMatchRatio = 0.1, windowLen = 10000,
matchWinLen = 10000, meanLogSeedLen = 1.7, sdLogSeedLen = 0.4,
seedPassRate = 0.78, sdTargetDist = 120, sameStrandProb = 0.5,
spikeWidth = 1500, betaShape1 = 0.5, betaShape2 = 0.5,
sameTarRegionProb = 0, chimMutRate = 0.005, noiseRate = 0.0015,
```

```
highNoiseRate = 0.08, highNoiseProb = 0.015, pairedEnd = TRUE,
prefix = "SimFFPE", threads = 1, localChimeric = TRUE,
distantChimeric = TRUE, normalReads = TRUE, overWrite = FALSE)
```

Arguments

sourceSeq	A DNAStringSet object of DNA sequences used for simulation. It can cover the entire reference genome or selected chromosomes or chromosome regions.
referencePath	Path to the reference genome.
PhredScoreProfile	A matrix representing the positional Phred score proportion. Each row of the matrix represents a position in the read (from begin to end), and each column the Phred quality score of base-calling error probabilities. The profile can be calculated from BAM file using the calcPhredScoreProfile function.
outFile	Output file path for the FASTQ file with simulated reads. Please include the name of the output file without extension, e.g. "/tmp/sim1".
coverage	Coverage of the simulation.
readLen	Read length of the simulation.
meanInsertLen	Mean insert length for the simulation (normally distributed).
sdInsertLen	Standard deviation of the insert length for simulation (normally distributed).
enzymeCut	Simulate enzymatic fragmentation if it is set to true, otherwise simulate random fragmentation.
chimericRatio	Proportion of artificial chimeric fragments (chimeric fragments / chimeric or normal fragments). Range: 0 to 1.
localMatchRatio	Proportion of adjacent ss-DNA combination (adjacent ss-DNA combination / adjacent or distant ss-DNA combination). Range: 0 to 1.
windowLen	The window length used in adjacent ss-DNA combination simulation. To simulate adjacent ss-DNA combinations, input DNA sequences are divided into small windows of equal size, and short complementary pairs are searched within the same window . Suggested range: 5000-20000. Unit: base pair (bp).
matchWinLen	The target window length used in distant ss-DNA simulation. To simulate distant ss-DNA combinations, the target sequences are searched in a random window. Suggested range: 5000-20000. Unit: base pair (bp).
meanLogSeedLen	Mean of log scaled seed length (bp). Seeds are used to search for complementary targets. The mapping of seed and target links two ss-DNA together, yielding artificial chimeric fragments. The seed length follows a log-normal distribution . See rlnorm for more details.
sdLogSeedLen	Standard deviation of log scaled seed length (bp).
seedPassRate	Proportion of seeds successfully forming chimeric fragments. Adjust this value when the percentage of chimeric reads in the output file is different from the parameter "chimericRatio".
sdTargetDist	Standard deviation of the normal distribution (mean = 0) used to simulate target selection probability. In adjacent ss-DNA combinations, when there are multiple targets for a seed, one target will be selected for combination. Target selection probability is simulated using the distance between seed and target. The smaller the distance, the larger the probability.

`sameStrandProb` Probability of seed and target from the same DNA strand (same strand ss-DNA combination / same or complementary strand ss-DNA combination). Only valid for adjacent ss-DNA combination. For paired end sequencing, the larger the probability, the greater the proportion of improperly paired reads with LL / RR pair orientation, and the smaller with RL pair orientation. Range: 0 to 1.

`spikeWidth` The width of chimeric read spike used to simulate distant ss-DNA combinations. In real FFPE samples, the chimeric reads formed by distant DNA combination are unevenly distributed along the chromosome. Some regions are enriched in these reads while some others are scarce. The length of these regions are of similar scale; therefore, a defined width is used for simulation. Suggested range: 1500-2000. Unit: base pair (bp).

`betaShape1` Shape parameter a of beta distribution used to model the unevenly distributed distant ss-DNA combinations. The number of seeds in each "spike" follows a "U" shaped beta distribution. Use this parameter to adjust the shape of the curve. See `rbeta` for more details. Range: 0-1.

`betaShape2` Shape parameter b of beta distribution used to model the unevenly distributed distant ss-DNA combinations. The number of seeds in each "spike" follows a "U" shaped beta distribution. Use this parameter to adjust the shape of the curve. See `rbeta` for more details. Range: 0-1.

`sameTarRegionProb`
Probability of neighboring seeds to search targets in same random region for distant ss-DNA combination simulation. The larger the value, the more the false positive translocation variants.

`chimMutRate` Mutation rate for each base in chimeric fragments. In the chimeric fragment formation process, biological-level errors might occur and lead to mutations on these artificial fragments. For all four basic types of nucleotides, the substitution probability is set equal. Range: 0-0.75.

`noiseRate` Noise rate for each base in reads. This is used for sequencing-level errors. The probability is set equal for all four basic types of nucleotides. Range: 0-0.75.

`highNoiseRate` A second noise rate for each base in reads. In some real sequencing data, some reads are much more noisy than others. This parameter can be used for this situation. Range: 0-0.75.

`highNoiseProb` Probability of reads to be simulated with highNoiseRate other than noiseRate. Range: 0-1.

`pairedEnd` Simulate paired end sequencing when set to true.

`prefix` Prefix for read names. When reads from different runs of simulation have to be merged, please make sure that they have different prefixes.

`threads` Number of threads used. Multi-threading can speed up the process.

`localChimeric` Generate reads from adjacent ss-DNA combinations if it is set to true. If it is set to false, skip this process.

`distantChimeric`
Generate reads from distant ss-DNA combinations if it is set to true. If it is set to false, skip this process.

`normalReads` Generate reads from normal fragments if it is set to true. If it is set to false, skip this process.

`overWrite` Overwrite the file if file with same output path exists and it is set to true. If file with same output path exists and it is set to false, reads will be appended to the existing file.

Details

The NGS (Next-Generation Sequencing) reads from FFPE (Formalin-Fixed Paraffin-Embedded) samples contain numerous artificial chimeric reads. These reads are derived from the combination of two single-stranded DNA (ss-DNA) fragments with short reverse complementary sequences. This function simulates these artificial reads as well as normal reads for FFPE samples on whole genome / several chromosomes / large regions. The combined ss-DNA may come from adjacent or distant regions. In the output fastq file these reads are distinguished by prefixes "localChimeric", "distantChimeric" and "Normal" in their names. The parameter PhredScoreProfile can be calculated by the function `calcPhredScoreProfile`. To simulate whole exome sequencing (WES) or targeted sequencing, please use the function `targetReadSimFFPE`.

Value

NULL. Reads will be written to the output FASTQ file.

Note

When fine-tuning is needed, simulate reads from certain areas / chromosomes instead of the entire genome to save the runtime. Please check the package vignette for the guidance of fine-tuning.

Author(s)

Lanying Wei <lanying.wei@uni-muenster.de>

See Also

`SimFFPE`, `calcPhredScoreProfile`, `targetReadSimFFPE`

Examples

```
PhredScoreProfilePath <- system.file("extdata", "PhredScoreProfile2.txt",
                                     package = "SimFFPE")
PhredScoreProfile <- as.matrix(read.table(PhredScoreProfilePath, skip = 1))
colnames(PhredScoreProfile) <- read.table(PhredScoreProfilePath,
                                          nrows = 1,
                                          colClasses = "character")

referencePath <- system.file("extdata", "example.fasta", package = "SimFFPE")
reference <- readDNAStringSet(referencePath)

## Simulate reads of the first three sequences of reference genome

sourceSeq <- reference[1:3]
outFile1 <- paste0(tempdir(), "/sim1")
readSimFFPE(sourceSeq, referencePath, PhredScoreProfile, outFile1,
            enzymeCut = FALSE, coverage=80, threads = 4)


## Simulate reads of defined regions on the first two sequences of reference
## genome

sourceSeq2 <- DNAStringSet(lapply(reference[1:2], function(x) x[1:10000]))
outFile2 <- paste0(tempdir(), "/sim2")
readSimFFPE(sourceSeq2, referencePath, PhredScoreProfile, outFile2,
            coverage = 80, enzymeCut = TRUE, threads = 1)
```

```
## Simulate reads of defined regions on the second and the third sequence of
## reference genome and merge them with existing reads (a different prefix is
## needed)

sourceSeq3 <- DNAStringSet(lapply(reference[2:3], function(x) x[1:10000]))
readSimFFPE(sourceSeq3, referencePath, PhredScoreProfile, outFile2,
            prefix = "simFFPE2", coverage = 80, enzymeCut = TRUE,
            threads = 1, overWrite = FALSE)
```

`targetReadSimFFPE` *Simulate noisy NGS reads of FFPE samples in exonic / targeted regions*

Description

NGS data from FFPE samples contain numerous artificial chimeric reads. These chimeric reads are formed through the combination of two single-stranded DNA (ss-DNA). This function simulates these artificial reads as well as normal reads for FFPE samples within defined regions.

Usage

```
targetReadSimFFPE(referencePath, PhredScoreProfile, targetRegions, outFile,
coverage, readLen = 150, meanInsertLen = 250, sdInsertLen = 80,
enzymeCut = FALSE, chimericRatio = 0.08, localMatchRatio = 0.1, padding = NA,
minGap = NA, windowLen = 10000, matchWinLen = 10000,
meanLogSeedLen = 1.7, sdLogSeedLen = 0.4, seedPassRate = 0.78, sdTargetDist=120,
sameStrandProb = 0.5, spikeWidth = 1500, betaShape1 = 0.5, betaShape2 = 0.5,
sameTarRegionProb = 0, chimMutRate = 0.005, noiseRate = 0.0015,
highNoiseRate = 0.08, highNoiseProb = 0.015, pairedEnd = TRUE,
prefix = "SimFFPE", threads = 1, localChimeric = TRUE,
distantChimeric = TRUE, normalReads = TRUE, overWrite = FALSE)
```

Arguments

`referencePath`	Path to the reference genome.
`PhredScoreProfile`	A matrix representing the positional Phred score proportion. Each row of the matrix represents a position in the read (from begin to end), and each column the Phred quality score of base-calling error probabilities. The profile can be calculated from BAM file using the `calcPhredScoreProfile` function.
`targetRegions`	A DataFrame or GenomicRanges object representing the exonic / targeted regions to simulate. If it is a DataFrame, the first column should be the chromosome, the second the start position and the third the end position. Please use one-based coordinate systems (the first base should be marked with 1 but not 0).
`outFile`	Output file path for the FASTQ file with simulated reads. Please include the name of the output file without extension, e.g. "/tmp/sim1".
`coverage`	Coverage of the simulation.
`readLen`	Read length of the simulation.

`meanInsertLen` Mean insert length for the simulation (normally distributed).

`sdInsertLen` Standard deviation of the insert length for simulation (normally distributed).

`enzymeCut` Simulate enzymatic fragmentation if it is set to true, otherwise simulate random fragmentation.

`chimericRatio` Proportion of artificial chimeric fragments (chimeric fragments / chimeric or normal fragments). Range: 0 to 1.

`localMatchRatio`
Proportion of adjacent ss-DNA combination (adjacent ss-DNA combination / adjacent or distant ss-DNA combination). Range: 0 to 1.

`padding` Length of padding of input target regions. The padding length will be added to both sides of target regions. If this value is not given, it will be assigned the value of input meanInsertLen divided by two. Range: natural numbers. Unit: base pair (bp).

`minGap` Minimal allowed length of gap between target regions. Regions with a gap smaller than this value will be merged. If this value is not given, the value of input readLen will be used. Range: natural numbers. Unit: base pair (bp).

`windowLen` The window length used in adjacent ss-DNA combination simulation. To simulate adjacent ss-DNA combinations, input DNA sequences are divided into small windows of equal size, and short complementary pairs are searched within the same window . Suggested range: 5000-20000. Unit: base pair (bp).

`matchWinLen` The target window length used in distant ss-DNA simulation. To simulate distant ss-DNA combinations, the target sequences are searched in a random window. Suggested range: 5000-20000. Unit: base pair (bp).

`meanLogSeedLen` Mean of log scaled seed length (bp). Seeds are used to search for complementary targets. The mapping of seed and target links two ss-DNA together, yielding artificial chimeric fragments. The seed length follows a log-normal distribution . See `rlnorm` for more details.

`sdLogSeedLen` Standard deviation of log scaled seed length (bp).

`seedPassRate` Proportion of seeds successfully forming chimeric fragments. Adjust this value when the percentage of chimeric reads in the output file is different from the parameter "chimericRatio".

`sdTargetDist` Standard deviation of the normal distribution (mean = 0) used to simulate target selection probability. In adjacent ss-DNA combinations, when there are multiple targets for a seed, one target will be selected for combination. Target selection probability is simulated using the distance between seed and target. The smaller the distance, the larger the probability.

`sameStrandProb` Probability of seed and target from the same DNA strand (same strand ss-DNA combination / same or complementary strand ss-DNA combination). Only valid for adjacent ss-DNA combination. For paired end sequencing, the larger the probability, the greater the proportion of improperly paired reads with LL / RR pair orientation, and the smaller with RL pair orientation. Range: 0 to 1.

`spikeWidth` The width of chimeric read spike used to simulate distant ss-DNA combinations. In real FFPE samples, the chimeric reads formed by distant DNA combination are unevenly distributed along the chromosome. Some regions are enriched in these reads while some others are scarce. The length of these regions are of similar scale; therefore, a defined width is used for simulation. Suggested range: 1500-2000. Unit: base pair (bp).

`betaShape1` Shape parameter a of beta distribution used to model the unevenly distributed distant ss-DNA combinations. The number of seeds in each "spike" follows a "U" shaped beta distribution. Use this parameter to adjust the shape of the curve. See `rbeta` for more details. Range: 0-1.

`betaShape2` Shape parameter b of beta distribution used to model the unevenly distributed distant ss-DNA combinations. The number of seeds in each "spike" follows a "U" shaped beta distribution. Use this parameter to adjust the shape of the curve. See `rbeta` for more details. Range: 0-1.

`sameTarRegionProb` Probability of neighboring seeds to search targets in same random region for distant ss-DNA combination simulation. The larger the value, the more the false positive translocation variants.

`chimMutRate` Mutation rate for each base in chimeric fragments. In the chimeric fragment formation process, biological-level errors might occur and lead to mutations on these artificial fragments. For all four basic types of nucleotides, the substitution probability is set equal. Range: 0-0.75.

`noiseRate` Noise rate for each base in reads. This is used for sequencing-level errors. The probability is set equal for all four basic types of nucleotides. Range: 0-0.75.

`highNoiseRate` A second noise rate for each base in reads. In some real sequencing data, some reads are much more noisy than others. This parameter can be used for this situation. Range: 0-0.75.

`highNoiseProb` Probability of reads to be simulated with highNoiseRate other than noiseRate. Range: 0-1.

`pairedEnd` Simulate paired end sequencing when set to true.

`prefix` Prefix for read names. When reads from different runs of simulation have to be merged, please make sure that they have different prefixes.

`threads` Number of threads used. Multi-threading can speed up the process.

`localChimeric` Generate reads from adjacent ss-DNA combinations if it is set to true. If it is set to false, skip this process.

`distantChimeric` Generate reads from distant ss-DNA combinations if it is set to true. If it is set to false, skip this process.

`normalReads` Generate reads from normal fragments if it is set to true. If it is set to false, skip this process.

`overWrite` Overwrite the file if file with same output path exists and it is set to true. If file with same output path exists and it is set to false, reads will be appended to the existing file.

Details

The NGS (Next-Generation Sequencing) reads from FFPE (Formalin-Fixed Paraffin-Embedded) samples contain numerous artificial chimeric reads. These reads are derived from the combination of two single-stranded DNA (ss-DNA) fragments with short reverse complementary sequences. This function simulates these artificial reads as well as normal reads for FFPE samples within defined regions. The combined ss-DNA may come from adjacent or distant regions. In the output fastq file these reads are distinguished by prefixes "localChimeric", "distantChimeric" and "Normal" in their names. The parameter PhredScoreProfile can be calculated by the function `calcPhredScoreProfile`. To simulate whole genome sequencing (WGS) or to simulate reads on several large regions / full chromosomes, please use the function `readSimFFPE`.

Value

NULL. Reads will be written to the output FASTQ file.

Note

When fine-tuning is needed, simulate reads from part of the regions instead of all the target regions to save the runtime. Please check the package vignette for the guidance of fine-tuning.

Author(s)

Lanying Wei <lanying.wei@uni-muenster.de>

See Also

SimFFPE, calcPhredScoreProfile, readSimFFPE

Examples

```
PhredScoreProfilePath <- system.file("extdata", "PhredScoreProfile1.txt",
                                     package = "SimFFPE")
PhredScoreProfile <- as.matrix(read.table(PhredScoreProfilePath, skip = 1))
colnames(PhredScoreProfile) <- read.table(PhredScoreProfilePath,
                                          nrows = 1,
                                          colClasses = "character")

referencePath <- system.file("extdata", "example.fasta", package = "SimFFPE")

regionPath <- system.file("extdata", "regionsSim.txt", package = "SimFFPE")
targetRegions <- read.table(regionPath)

outFile <- paste0(tempdir(), "/sim3")
targetReadSimFFPE(referencePath, PhredScoreProfile, targetRegions, outFile,
                  coverage = 80, readLen = 100, meanInsertLen=180,
                  sdInsertLen=50, enzymeCut = FALSE)
```

Index