Package 'MicroSEC'

December 2, 2020

Title Sequence Error Filtering Pipeline for FFPE Samples

Version 1.1.3

Description Clinical sequencing of tumor is usually performed on formalin-fixed and paraffin-embedded (FFPE) samples and have many sequencing errors. We found that the majority of these errors are detected in chimeric read caused by single-strand DNA with microhomology. Our filtering pipeline focuses on the uneven distribution of the artifacts in each read and removes such errors in FFPE samples without over-eliminating the true mutations detected in fresh frozen samples.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Depends R (>= 3.4.0)

RoxygenNote 7.1.1

Imports tidyr, openxlsx, data.table, R.utils, stringr, magrittr, dplyr, gtools, Biostrings, GenomicAlignments, Rsamtools, GenomeInfoDb, BiocGenerics

Suggests BSgenome.Hsapiens.UCSC.hg38, BSgenome.Hsapiens.UCSC.hg19, BSgenome.Mmusculus.UCSC.mm10, knitr, rmarkdown

VignetteBuilder knitr

URL https://github.com/MANO-B/MicroSEC/

BugReports https://github.com/MANO-B/MicroSEC/issues

NeedsCompilation no

Author Masachika Ikegami [aut, cre]

Maintainer Masachika Ikegami <ikegamitky@gmail.com>

Repository CRAN

Date/Publication 2020-12-02 10:30:07 UTC

R topics documented:

`exampleBAM` *An example BAM file.*

Description

A BAM file containing the information of eight mutations.

Usage

```
exampleBAM
```

Format

A list with 7 factors, each contains 184 variables:

rname chromosome of the read

qname read ID list

seq sequence of the read, in DNAString

strand strand of the read

cigar CIGAR sequence of the read

qual Phred quality of the read

pos starting position of the read ...

---

exampleMutation *An example mutation file.*

Description

A dataset containing the information of eight mutations.

Usage

```
exampleMutation
```

Format

A list with 14 variables:

Sample sample name

Gene altered gene

HGVS.c base change

HGVS.p protein change

Mut_type mutation type

Total_QV>=20 total reads

%Alt altered base ratio

Chr altered chromosome

Pos altered position

Ref reference base

Alt altered base

SimpleRepeat_TRF mutation locating repeat sequence

Neighborhood_sequence neighborhood sequence

Transition base change type ...

exampleMutCall *An example mutated read ID file.*

Description

A dataset containing the information of mutated read.

Usage

```
exampleMutCall
```

Format

A list with 7 factors, each contains 184 variables:

Chr altered chromosome

Pos altered position

Ref reference base

Alt altered base

Mut_ID mutated read ID

Mut mutation detail ...

fun_analysis *Analyzing function.*

Description

This function analyzes the filtering results.

Usage

```
fun_analysis(
  msec,
  mut_depth,
  short_homology_search_length,
  min_homology_search,
  threshold_p,
  threshold_hairpin_ratio,
  threshold_soft_clip_ratio,
  threshold_short_length,
  threshold_distant_homology,
  threshold_low_quality_rate,
  homopolymer_length
)
```

Arguments

msec	Mutation filtering information.
mut_depth	Mutation coverage data.
short_homology_search_length	Small sequence for homology search.
min_homology_search	The sequence length for homology search.
threshold_p	The largest p value of significant errors.
threshold_hairpin_ratio	The smallest hairpin read ratio.
threshold_soft_clip_ratio	The smallest rate of significantly soft-clipped reads.
threshold_short_length	Reads shorter than that are analyzed.
threshold_distant_homology	The smallest rate of reads from other regions.
threshold_low_quality_rate	The smallest rate of low quality bases.
homopolymer_length	The smallest length of homopolymers.

Value

msec

Examples

```
fun_analysis(msec = msec_summarized,
             mut_depth = mut_depth_checked,
             short_homology_search_length = 4,
             min_homology_search = 40,
             threshold_p = 10 ^ (-6),
             threshold_hairpin_ratio = 0.50,
             threshold_soft_clip_ratio = 0.90,
             threshold_short_length = 0.8,
             threshold_distant_homology = 0.2,
             threshold_low_quality_rate = 0.1,
             homopolymer_length = 15)
```

fun_convert *Mutation data file converter*

Description

This function attempts to convert the mutation information file.

Usage

```
fun_convert(mutation_file, organism)
```

Arguments

mutation_file	Path of the mutation information file.
organism	Human or Mouse genome.

Value

df_mutation

Examples

```
fun_convert(
  system.file("extdata", "convert_test.xlsx", package = "MicroSEC"),
  "hg19"
)
```

fun_hairpin_check *Hairpin-structure sequence check function*

Description

This function attempts to find hairpin structure sequences.

Usage

```
fun_hairpin_check(hairpin_seq_tmp, ref_seq, hairpin_length, hair)
```

Arguments

hairpin_seq_tmp	The sequence to be checked.
ref_seq	Reference sequence around the mutation.
hairpin_length	The temporal length of hairpin sequences.
hair	The length of sequences to be checked.

Value

list(hairpin_length, whether hairpin sequences exist or not)

`fun_hairpin_trimming` *Adapter sequence trimming function*

Description

This function attempts to find and cut the adapter sequences in the ends.

Usage

```
fun_hairpin_trimming(hairpin_seq, mut_read_strand, adapter_1, adapter_2)
```

Arguments

`hairpin_seq`	The sequence to be trimmed.
`mut_read_strand`	The strand of the sequence, "+" or "-".
`adapter_1`	The Read 1 adapter sequence of the library.
`adapter_2`	The Read 2 adapter sequence of the library.

Value

Adapter-trimmed hairpin_seq

`fun_homology` *Homology check function.*

Description

This function attempts to search the homologous regions.

Usage

```
fun_homology(
  msec,
  df_distant,
  min_homology_search,
  ref_genome,
  chr_no,
  progress_bar
)
```

Arguments

msec	Mutation filtering information.
df_distant	Sequences to be checked.
min_homology_search	Minimum length to define "homologous".
ref_genome	Reference genome.
chr_no	Chromosome number.
progress_bar	"Y": You can see the progress visually.

Value

msec

Examples

```
fun_homology(msec = msec_read_checked,
             df_distant = homology_searched,
             min_homology_search = 40,
             ref_genome = BSgenome.Hsapiens.UCSC.hg38::
                          BSgenome.Hsapiens.UCSC.hg38,
             chr_no = 24,
             progress_bar = "Y")
```

---

fun_insert_length *Insert length checker for BAM file*

Description

This function attempts to check the insert length in BAM file.

Usage

```
fun_insert_length(bam_file)
```

Arguments

bam_file	Path of the BAM file.

Value

df_bam

Examples

```
fun_insert_length(
  system.file("extdata", "test.bam", package = "MicroSEC")
)
```

fun_load_bam *BAM file loader*

Description

This function attempts to load the BAM file.

Usage

```
fun_load_bam(bam_file)
```

Arguments

bam_file	Path of the BAM file.

Value

df_bam

Examples

```
fun_load_bam(
  system.file("extdata", "test.bam", package = "MicroSEC")
)
```

fun_load_chr_no *Chromosome number loading function.*

Description

This function attempts to load the chromosome number.

Usage

```
fun_load_chr_no(organism)
```

Arguments

organism	Human or Mouse genome.

Value

chr_no

Examples

```
fun_load_chr_no("Human")
```

fun_load_genome *Genome loading function.*

Description

This function attempts to load the appropriate genome.

Usage

```
fun_load_genome(organism)
```

Arguments

organism	Human or Mouse genome.

Value

ref_genome

Examples

```
fun_load_genome("Human")
```

fun_load_id *Mutation-supporting read ID files loader*

Description

This function attempts to load the read ID information files.

Usage

```
fun_load_id(read_list)
```

Arguments

read_list	Path of the read ID information directory.

Value

df_mut_call

Examples

```
fun_load_id(
  system.file("extdata", package = "MicroSEC")
)
```

fun_load_mutation *Mutation data file loader*

Description

This function attempts to load the mutation information file.

Usage

```
fun_load_mutation(mutation_file, sample_name)
```

Arguments

mutation_file	Path of the mutation information file.
sample_name	Sample name.

Value

df_mutation

Examples

```
fun_load_mutation(
  system.file("extdata", "test_mutation.xlsx", package = "MicroSEC"),
 "H15-11943-1-T_TDv3"
)
```

fun_load_mutation_gz *Mutation data file loader*

Description

This function attempts to load the mutation information file.

Usage

```
fun_load_mutation_gz(mutation_file)
```

Arguments

mutation_file	Path of the mutation information file.

Value

df_mutation

Examples

```
fun_load_mutation_gz(
  system.file("extdata", "test_mutation.tsv", package = "MicroSEC")
)
```

`fun_read_check` *Read check function.*

Description

This function attempts to check the mutation profile in each read.

Usage

```
fun_read_check(
  df_mutation,
  df_bam,
  df_mut_call,
  ref_genome,
  sample_name,
  read_length,
  adapter_1,
  adapter_2,
  short_homology_search_length,
  progress_bar
)
```

Arguments

`df_mutation`	Mutation information.
`df_bam`	Data from the BAM file.
`df_mut_call`	Read ID list.
`ref_genome`	Reference genome for the data.
`sample_name`	Sample name (character)
`read_length`	The read length in the sequence.
`adapter_1`	The Read 1 adapter sequence of the library.
`adapter_2`	The Read 2 adapter sequence of the library.
`short_homology_search_length`	Small sequence for homology search.
`progress_bar`	"Y": You can see the progress visually.

Value

list(msec, homology_search)

Examples

```
fun_read_check(df_mutation = exampleMutation,
               df_bam =  exampleBAM,
               df_mut_call = exampleMutCall,
               ref_genome = BSgenome.Hsapiens.UCSC.hg38::
                             BSgenome.Hsapiens.UCSC.hg38,
               sample_name = "H15-11943-1-T_TDv3",
               read_length = 151,
               adapter_1 = "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA",
               adapter_2 = "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT",
               short_homology_search_length = 4,
               progress_bar = "Y")
```

`fun_repeat_check` *Repeat check function.*

Description

This function attempts to check the repetitive sequence around the mutation.

Usage

```
fun_repeat_check(rep_a, rep_b, ref_seq, ref_width, del)
```

Arguments

`rep_a`	The shorter sequence of Ref and Alt.
`rep_b`	The longer sequence of Ref and Alt.
`ref_seq`	Reference sequence around the mutation.
`ref_width`	Search length for ref_seq.
`del`	Insertion: 0, Deletion: 1

Value

list(pre_rep_status, post_rep_status, homopolymer_status)

fun_save *Save function.*

Description

This function attempts to save the filtering results.

Usage

```
fun_save(msec, sample_info, wd)
```

Arguments

msec	Mutation filtering information.
sample_info	The sample name.
wd	The directory to save.

Examples

```
fun_save(msec_analyzed, "test_data", ".")
```

fun_save_gz *Save function.*

Description

This function attempts to save the filtering results.

Usage

```
fun_save_gz(msec, output)
```

Arguments

msec	Mutation filtering information.
output	output file name (full path).

Examples

```
fun_save_gz(msec_analyzed, "./test_data.tsv")
```

fun_setting *Mutated position search function.*

Description

This function attempts to find the mutated bases in each read.

Usage

```
fun_setting(pre, post, neighbor_seq, neighbor_length, alt_length)
```

Arguments

pre	The 5' side bases of the sequence for searching.
post	The 3' side bases of the sequence for searching.
neighbor_seq	Short reference sequence around the mutation.
neighbor_length	The length from the mutation to the ends of the short reference sequence.
alt_length	The length of altered bases.

Value

list(pre_search_length, post_search_length, peri_seq_1, peri_seq_2)

fun_summary *Summarizing function.*

Description

This function summarizes the filtering results.

Usage

```
fun_summary(msec)
```

Arguments

msec	Mutation filtering information.

Value

msec

Examples

```
fun_summary(msec_homology_searched)
```

`fun_support` *Supporting length calculation function.*

Description

This function attempts to calculate supporting lengths of a read.

Usage

```
fun_support(
  df_cigar,
  df_seq,
  mut_read_strand,
  adapter_1,
  adapter_2,
  mut_position,
  alt_length,
  indel_status
)
```

Arguments

`df_cigar`	The CIGAR data of the read.
`df_seq`	The sequence to be checked.
`mut_read_strand`	The strand of the sequence, "+" or "-".
`adapter_1`	The Read 1 adapter sequence of the library.
`adapter_2`	The Read 2 adapter sequence of the library.
`mut_position`	The mutation position in the read.
`alt_length`	The length of altered bases.
`indel_status`	The mutation is indel or not.

Value

list(pre_support_length, post_support_length, soft_clipped_read)

fun_zero *Devide function without 0/0 errors*

Description

This function attempts to devide without 0/0 errors.

Usage

```
fun_zero(a, b)
```

Arguments

a, b	Integers

Value

a devided by b

homology_searched *An example sequence infromation file.*

Description

A dataset containing the information of reads for homology search.

Usage

```
homology_searched
```

Format

A list with 7 factors, each contains 460 variables:

sample_name sample name

Chr altered chromosome

Pos altered position

Ref reference base

Alt altered base

Direction 5' (pre) or 3' (post) sequence of the mutated base

Seq sequence for homology search ...

`msec_analyzed` *An example mutation file.*

Description

A dataset containing the information of eight mutations processed by the fun_homology function.

Usage

```
msec_analyzed
```

Format

A list with 38 variables:

Sample sample name

Gene altered gene

HGVS.c base change

HGVS.p protein change

Mut_type mutation type

Total_QV>=20 total reads

%Alt altered base ratio

Chr altered chromosome

Pos altered position

Ref reference base

Alt altered base

SimpleRepeat_TRF mutation locating repeat sequence

Neighborhood_sequence neighborhood sequence

Transition base change type

read_length read length

total_read number of mutation supporting reads

soft_clipped_read number of soft-clipped reads

flag_hairpin number of reads produced by hairpin structure

pre_support_length muximum 5'-supporting length

post_support_length muximum 3'-supporting length

short_support_length minimum supporting length

low_quality_base_rate_under_q18 low quality base rate

distant_homology_rate rate of reads derived from homologous regions

prob_filter_1 possibility of short-supporting length

prob_filter_3_pre possibility of 5'-supporting length

prob_filter_3_post possibility of 3'-supporting length

filter_1_mutation_intra_hairpin_loop filter 1

filter_2_hairpin_structure filter 2

filter_3_microhomology_induced_mutation filter 3

filter_4_soft_clipping filter 4

filter_5_highly_homologous_region filter 5

filter_6_simple_repeat filter 6

filter_7_c_to_t_artifact filter 7

filter_8_mutation_at_homopolymer filter 8

msec_filter_1234 any of filter 1-4

msec_filter_12345 any of filter 1-5

msec_filter_all any of filter 1-8

comment comment ...

`msec_homology_searched`

An example mutation file.

Description

A dataset containing the information of eight mutations processed by the fun_homology function.

Usage

```
msec_homology_searched
```

Format

A list with 36 variables:

Sample sample name

Gene altered gene

HGVS.c base change

HGVS.p protein change

Mut_type mutation type

Total_QV>=20 total reads

%Alt altered base ratio

Chr altered chromosome

Pos altered position

Ref reference base

Alt altered base

SimpleRepeat_TRF mutation locating repeat sequence

Neighborhood_sequence neighborhood sequence

Transition base change type

read_length read length

mut_type mutation type

alt_length length of the mutated bases

total_read number of mutation supporting reads

soft_clipped_read number of soft-clipped reads

flag_hairpin number of reads produced by hairpin structure

hairpin_length maximum length of palindromic sequences

pre_support_length muximum 5'-supporting length

post_support_length muximum 3'-supporting length

short_support_length minimum supporting length

pre_minimum_length minimum 5'-supporting length

post_minimum_length minimum 3'-supporting length

low_quality_base_rate_under_q18 low quality base rate

pre_rep_status 5'-repeat sequence length

post_rep_status 3'-repeat sequence length

homopolymer_status homopolymer sequence length

indel_status whether the mutation is indel or not

indel_length length of indel mutation

distant_homology number of reads derived from homologous regions

penalty_pre 5'-penalty score by the mapper

penalty_post 3'-penalty score by the mapper

caution comment ...

`msec_read_checked` *An example mutation file.*

Description

A dataset containing the information of eight mutations processed by the fun_read_check function.

Usage

```
msec_read_checked
```

Format

A list with 36 variables:

Sample sample name
Gene altered gene
HGVS.c base change
HGVS.p protein change
Mut_type mutation type
Total_QV>=20 total reads
%Alt altered base ratio
Chr altered chromosome
Pos altered position
Ref reference base
Alt altered base
SimpleRepeat_TRF mutation locating repeat sequence
Neighborhood_sequence neighborhood sequence
Transition base change type
read_length read length
mut_type mutation type
alt_length length of the mutated bases
total_read number of mutation supporting reads
soft_clipped_read number of soft-clipped reads
flag_hairpin number of reads produced by hairpin structure
hairpin_length maximum length of palindromic sequences
pre_support_length muximum 5'-supporting length
post_support_length muximum 3'-supporting length
short_support_length minimum supporting length
pre_minimum_length minimum 5'-supporting length
post_minimum_length minimum 3'-supporting length
low_quality_base_rate_under_q18 low quality base rate
pre_rep_status 5'-repeat sequence length
post_rep_status 3'-repeat sequence length
homopolymer_status homopolymer sequence length
indel_status whether the mutation is indel or not
indel_length length of indel mutation
distant_homology number of reads derived from homologous regions
penalty_pre 5'-penalty score by the mapper
penalty_post 3'-penalty score by the mapper
caution comment ...

`msec_summarized` *An example mutation file.*

Description

A dataset containing the information of eight mutations processed by the fun_homology function.

Usage

```
msec_summarized
```

Format

A list with 53 variables:

Sample sample name

Gene altered gene

HGVS.c base change

HGVS.p protein change

Mut_type mutation type

Total_QV>=20 total reads

%Alt altered base ratio

Chr altered chromosome

Pos altered position

Ref reference base

Alt altered base

SimpleRepeat_TRF mutation locating repeat sequence

Neighborhood_sequence neighborhood sequence

Transition base change type

read_length read length

mut_type mutation type

alt_length length of the mutated bases

total_read number of mutation supporting reads

soft_clipped_read number of soft-clipped reads

flag_hairpin number of reads produced by hairpin structure

hairpin_length maximum length of palindromic sequences

pre_support_length muximum 5'-supporting length

post_support_length muximum 3'-supporting length

short_support_length minimum supporting length

pre_minimum_length minimum 5'-supporting length

post_minimum_length minimum 3'-supporting length

low_quality_base_rate_under_q18 low quality base rate

pre_rep_status 5'-repeat sequence length

post_rep_status 3'-repeat sequence length

homopolymer_status homopolymer sequence length

indel_status whether the mutation is indel or not

indel_length length of indel mutation

distant_homology number of reads derived from homologous regions

penalty_pre 5'-penalty score by the mapper

penalty_post 3'-penalty score by the mapper

caution comment

distant_homology_rate rate of reads derived from homologous regions

pre_minimum_length_adj adjusted pre_minimum_length

post_minimum_length_adj adjusted pre_minimum_length

pre_support_length_adj adjusted pre_minimum_length

post_support_length_adj adjusted pre_minimum_length

shortest_support_length_adj the shortest short_support_length

minimum_length_1 theoretically minimum 5'-supporting length

minimum_length_2 theoretically minimum 3'-supporting length

minimum_length theoretically minimum supporting length

short_support_length_adj adjusted short_support_length

altered_length substituted/inserted length

half_length half of the read length

short_support_length_total range of short_support_length

pre_support_length_total range of pre_support_length

post_support_length_total range of post_support_length

half_length_total range of possible short_support_length

total_length_total range of possible supporting length ...

`mut_depth_checked` *An example sequence infromation file.*

Description

A dataset containing the information of reads for homology search.

Usage

```
mut_depth_checked
```

Format

A list with 7 factors, each contains 460 variables:

Zero zero vector

Depth0 number of reads whose 5'-supporting length is <= 0

Depth1 number of reads whose 5'-supporting length is <= 1

Depth2 number of reads whose 5'-supporting length is <= 2

Depth3 number of reads whose 5'-supporting length is <= 3 ...

Index