Package 'fRagmentomics'

May 28, 2025

Title Compute Fragments Statistics Of Sequencing Experiments

Version 0.2.0

Description A user-friendly R package that enables the characterization of each cfDNA fragment overlapping one or multiple mutations of interest, starting from a sequencing file containing aligned reads (BAM file). fRagmentomics supports multiple mutation input formats (e.g., VCF, TSV, or string "chr:pos:ref:alt" representation), accommodates one-based and zero-based genomic conventions, handles mutation representation ambiguities, and accepts any reference file and species in FASTA format. For each cfDNA fragment, fRagmentomics outputs its size, its 3' and 5' sequences, and its mutational status.

License GPL (>= 3)

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Suggests covr,
testthat (>= 3.0.0)

Config/testthat/edition 3

Imports doParallel,
foreach,
parallel,
readr,
stringr,
Rsamtools,
GenomicRanges,
Biostrings,
IRanges,
GenomeInfoDb

Contents

analyze_fragments 2
fRagmentomics 3
process_fragment 4

Index **6**

analyze_fragments *Analyze fragments*

Description

Analyze fragments

Usage

```
analyze_fragments(
  mut,
  bam,
  fasta,
  sample_id = NA,
  neg_offset_mate_search = -1000,
  pos_offset_mate_search = 1000,
  one_based = TRUE,
  flag_keep = 3,
  flag_remove = 2304,
  report_tlen = FALSE,
  report_softclip = FALSE,
  report_5p_3p_bases_fragment = 5,
  cigar_free_mode = FALSE,
  tmp_folder = tempdir(),
  output_file = NA,
  n_cores = 1
)
```

Arguments

mut	Path to a .vcf or .tsv file or string representation chr:pos:ref:alt of a mutation.
bam	Path to a BAM file.
fasta	Path to the FASTA file for the reference sequence used for generating the BAM file.
sample_id	Sample identifier.
neg_offset_mate_search	Integer. Use in read_bam. Represents the number of nucleotides to extend upstream (negative direction) from the position of interest when querying the BAM file with Rsamtools. his extension ensures that paired reads are retrieved, even if only one mate overlaps the queried position.
pos_offset_mate_search	Integer. Use in read_bam.
one_based	Boolean. TRUE if fasta is in one based. False if in 0 based.
flag_keep	Character vector. Use in read_bam. Represents the SAM flags that should be kept when filtering alignments.
flag_remove	Character vector. Use in read_bam. Represents the SAM flags that should be excluded when filtering alignments.
report_tlen	Boolean. Whether to include the TLEN (template length) information in the output.

`report_softclip`
: Boolean. Whether to include the number of soft-clipped bases at the fragment extremities in the output.

`report_5p_3p_bases_fragment`
: Integer. Whether to include N fragment extremity bases in the output.

`cigar_free_mode`
: Boolean. If activated, the information from the CIGAR is disregarded when determining the mutation status of a read. Instead the mutation status is determined by comparing the sequence of the read to the sequence of the wild-type reference and the mutated reference. Activating this option may lead to discordant genotyping of reads compared to the information provided by the CIGAR for indels. On the other hand, when activated, it may rescue mutated genotypes for indel that would be missed in cases where the representation of the indel in the CIGAR does not match the norm of bcftools of the mutation being analyzed.

`tmp_folder`
: Character vector for the folder temporary path.

`output_file`
: Character vector for the output file path. Mandatory.

`n_cores`
: Number of cores for parallel computation.

Value

A dataframe containing extracted fragment-level information.

`fRagmentomics` *fRagmentomics: Characterization of cfDNA Fragments*

Description

A user-friendly R package that enables the characterization of each cfDNA fragment overlapping one or multiple mutations of interest, starting from a sequencing file containing aligned reads (BAM file). fRagmentomics supports multiple mutation input formats (e.g., VCF, TSV, or string "chr:pos:ref:alt" representation), accommodates one-based and zero-based genomic conventions, handles mutation representation ambiguities, and accepts any reference file and species in FASTA format. For each cfDNA fragment, fRagmentomics outputs its size, its 3' and 5' sequences, and its mutational status.

Author(s)

Killian Maudet killian.maudet@gustaveroussy.fr aut, cre

Yoann Pradat yoann.pradat@gustaveroussy.fr aut, cre

Juliette Samaniego juliette.samaniego@gustaveroussy.fr aut

Elsa Bernard elsa.bernard@gustaveroussy.fr aut

Examples

```
# Example usage of fRagmentomics
# Load the package
library(fRagmentomics)

# Example function call
# result <- analyze_fragments(bam_file = "path/to/your/file.bam",
#                             mutation_file = "path/to/your/mutations.vcf")
```

process_fragment *Process a single sequencing fragment Extracts and processes relevant data from a sequencing fragment.*

Description

Process a single sequencing fragment Extracts and processes relevant data from a sequencing fragment.

Usage

```
process_fragment(
  df_sam,
  fragment_name,
  sample_id,
  chr,
  pos,
  ref,
  alt,
  report_tlen,
  report_softclip,
  report_5p_3p_bases_fragment,
  cigar_free_mode,
  fasta_fafile
)
```

Arguments

df_sam	A dataframe containing sequencing reads.
fragment_name	Name of the fragment (paired-end reads).
sample_id	Sample identifier.
chr	Character vector representing the chromosome of interest.
pos	Numeric value representing the Genomic position of interest.
ref	Character vector representing reference base(s).
alt	Character vector representing alternative base(s).
report_tlen	Boolean. Whether to include the TLEN (template length) information in the output.
report_softclip	Boolean. Whether to include the number of soft-clipped bases at the fragment extremities in the output.
report_5p_3p_bases_fragment	Integer. Whether to include N fragment extremity bases in the output.
cigar_free_mode	Boolean. If activated, the information from the CIGAR is disregarded when determining the mutation status of a read. Instead the mutation status is determined by comparing the sequence of the read to the sequence of the wild-type reference and the mutated reference. Activating this option may lead to discordant genotyping of reads compared to the information provided by the CIGAR

for indels. On the other hand, when activated, it may rescue mutated genotypes for indel that would be missed in cases where the representation of the indel in the CIGAR does not match the norm of bcftools of the mutation being analyzed.

`fasta_fafile` An open connection to an object of class FaFile

Value

A dataframe with the processed fragment information.

Index

analyze_fragments, 2
aut, *3*
aut, cre, *3*

fRagmentomics, 3

process_fragment, 4