Introduction to ALKfusiondiscover

2022-07-18

Introduction

This is an introduction to the **ALKfusiondiscover** package which can be used to evaluate whether EML4-ALK is present in blood plasma.

This package was created in order to increase the sensitivity of EML4-ALK detection from commercially available NGS products such the AVENIO (Roche) pipeline.

Paired-end sequencing of cfDNA generated BAM files can be used as input to discover EML4-ALK variants. This package was developed using position deduplicated BAM files generated with the AVENIO Oncology Analysis Software. These files are made using the AVENIO ctDNA surveillance kit and Illumina Nextseq 500 sequencing. This is a targeted hybridization NGS approach and includes ALK-specific but not EML4-specific probes.

The package includes six functions.

The output of first function, `EML4_ALK_detection()`, is used to determine whether EML4-ALK is detected and serves as input for the next four exploratory functions characterizing the EML4-ALK variant. The last function `EML4_ALK_analysis()` combines the output of the exploratory functions.

To serve as examples, this package includes BAM files representing the EML4-ALK positive cell line H3122 and the EML4-ALK negative cell line, HCC827.

Installation

Use **devtools** to install **ALKfusiondiscover**.

```
if (!require(devtools)) install.packages('devtools')
library(devtools)

install_github("CTrierMaansson/ALKfusiondiscover")
library(ALKfusiondiscover)
```

Package data

BAM files from the cell lines, H3122 and HCC827, are included in the package and can be used as examples to explore the functions.

```
H3122_bam <- system.file("extdata", "H3122_EML4.bam", package = "ALKfusiondiscover")
HCC827_bam <-  system.file("extdata", "HCC827_EML4.bam", package = "ALKfusiondiscover")
```

`EML4_ALK_detection()`

This function looks for EML4-ALK mate pair reads in the BAM file.

Input:

`file`

The name of the file which the data are to be read from.

`genome`

`character` representing the reference genome. Can be either “hg38” or “hg19”. Default = “hg38”.

`mates`

`interger`, the minimum number EML4-ALK mate pairs needed to be detected in order to call a variant. Default = 2.

Output:

If EML4-ALK is detected a `data.frame` with soft-clipped reads representing EML4-ALK is returned. Otherwise “No EML4-ALK was detected” is returned.

Examples:

```
head(EML4_ALK_detection(file = H3122_bam, genome = "hg38", mates = 2))
#>                                                                                          sequences
#> 1 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 2 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 3 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 4 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 5 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 6 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#>       mate position cigar
#> 1 29223691 42299657 94M2S
#> 2 29223375 42299657 94M2S
#> 3 29223479 42299657 94M2S
#> 4 29223686 42299657 94M2S
#> 5 29223636 42299657 94M2S
#> 6 29223687 42299657 94M2S
EML4_ALK_detection(file = HCC827_bam, genome = "hg38", mates = 2)
#> [1] "No EML4-ALK was detected"
```

`EML4_sequence()`

This function identifies the basepairs leading up to the EML4 breakpoint.

Input:

`reads`

`data.frame` returned by EML4_ALK_detection().

`basepairs`

`integer`, number of basepairs identified from the EML4-ALK fusion. Default = 20.

Output:

If EML4-ALK is detected, returns a `table` of identified EML4 basepairs with the number of corresponding reads for each sequence. Otherwise “No EML4-ALK was detected” is returned.

Examples:

```
EML4_sequence(EML4_ALK_detection(file = H3122_bam, genome = "hg38", mates = 2), basepairs = 20)
#> EML4_seq
#> CCAGGCTGGAGTGCAGTGGT GGAGTGCAGTGGTGTGATTT TCAGGCTGGAGTGCAGTGGT
#>                  201                    1                    1
EML4_sequence(EML4_ALK_detection(file = HCC827_bam, genome = "hg38", mates = 2), basepairs = 20)
#> [1] "No EML4-ALK was detected"
```

`ALK_sequence()`

This function identifies the basepairs following the ALK breakpoint.

Input:

`reads`

`data.frame` returned by EML4_ALK_detection().

`basepairs`

`integer`, number of basepairs identified from the EML4-ALK fusion. Default = 20.

Output:

If EML4-ALK is detected, returns a `table` of identified ALK basepairs with the number of corresponding reads for each sequence. Otherwise “No EML4-ALK was detected” is returned.

Examples:

```
ALK_sequence(EML4_ALK_detection(file = H3122_bam, genome = "hg38", mates = 2), basepairs = 20)
#> ALK_seq
#> CAGAATTTTAGCTTTGCAAT      CGGATTTTTAGCTTT CGGATTTTTAGCTTTTCATT
#>                    1                    1                    2
#> CGGAATTTTAGCTTTGCATT                   CT                  CTG
#>                    1                    8                    3
#>                 CTGA                CTGAT             CTGATTTT
#>                   11                   16                    5
#>            CTGATTTTT           CTGATTTTTA          CTGATTTTTAG
#>                    6                    3                    3
#> CTGATTTTTAGATTTGCATT         CTGATTTTTAGC        CTGATTTTTAGCT
#>                    1                   14                   10
#>       CTGATTTTTAGCTT      CTGATTTTTAGCTTT     CTGATTTTTAGCTTTG
#>                   10                    3                    4
#>    CTGATTTTTAGCTTTGC   CTGATTTTTAGCTTTGCA  CTGATTTTTAGCTTTGCAT
#>                    7                    8                    1
#> CTGATTTTTAGCTTTGCATT CTGATTTTTAGCTTTGCAAT     CTGATTTTTAGCTTTT
#>                   71                    1                    1
#> CTGATTTTTAGCTTTTCATA          CTGATTTTTAT     CTGATTTTTATCTTTG
#>                    1                    1                    2
#> CTGATTTTTATCTTTGCATT CTGATTTTTATCTTTTGATT CTGTGTTTTAGATTTGCATT
#>                    2                    1                    1
#> CTGTTTTTTATCTTTGCAAT                CTGAA CTTATTTTTATCTTTGCATT
#>                    1                    1                    1
#>            TTAGCTTTG
#>                    1
ALK_sequence(EML4_ALK_detection(file = HCC827_bam, genome = "hg38", mates = 2), basepairs = 20)
#> [1] "No EML4-ALK was detected"
```

`break_position()`

This function identifies the genomic position in EML4 where the breakpoint has happened.

Input:

`reads`

`data.frame` returned by EML4_ALK_detection().

Output:

If EML4-ALK is detected, returns a `table` of genomic positions with the number of corresponding reads for each sequence. Otherwise "No EML4-ALK was detected" is returned.

Examples:

```
break_position(EML4_ALK_detection(file = H3122_bam, genome = "hg38", mates = 2))
#> break_pos
#> 42299750 42299757
#>      202        1
break_position(EML4_ALK_detection(file = HCC827_bam, genome = "hg38", mates = 2))
#> [1] "No EML4-ALK was detected"
```

`break_position_depth()`

This function identifies the read depth at the basepair before the breakpoint in EML4.

Input:

`file`

The name of the file which the data are to be read from.

`reads`

`data.frame` returned by EML4_ALK_detection().

Output:

If EML4-ALK is detected a single `integer` corresponding to the read depth at the breakpoint is returned. Otherwise “No EML4-ALK was detected” is returned

Examples:

```
break_position_depth(H3122_bam, EML4_ALK_detection(file = H3122_bam, genome = "hg38", mates = 2))
#> [1] 251
break_position_depth(HCC827_bam, EML4_ALK_detection(file = HCC827_bam, genome = "hg38", mates = 2))
#> [1] "No EML4-ALK was detected"
```

`EML4_ALK_analysis()`

This functions collects the results from the other functions of the package.

Input:

`file`

The name of the file which the data are to be read from.

`genome`

`character` representing the reference genome. Can be either “hg38” or “hg19”. Default = “hg38”.

`mates`

`interger`, the minimum number EML4-ALK mate pairs needed to be detected in order to call a variant. Default = 2.

`basepairs`

`integer`, number of basepairs identified from the EML4-ALK fusion. Default = 20.

Output:

A `list` object with clipped_reads corresponding to `EML4_ALK_detection()`, last_EML4 corresponding to `EML4_sequence()`, first_ALK corresponding to `ALK_sequence()`, breakpoint corresponding to `break_position()`, and read_depth corresponding to `break_position_depth()`.

Examples:

```r
H3122_results <- EML4_ALK_analysis(file = H3122_bam, genome = "hg38", mates = 2, basepairs = 20)
HCC827_results <- EML4_ALK_analysis(file = HCC827_bam, genome = "hg38", mates = 2, basepairs = 20)
```

```r
head(H3122_results$clipped_reads)
#>                                                                                          sequences
#> 1 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 2 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 3 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 4 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 5 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 6 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#>       mate position cigar
#> 1 29223691 42299657 94M2S
#> 2 29223375 42299657 94M2S
#> 3 29223479 42299657 94M2S
#> 4 29223686 42299657 94M2S
#> 5 29223636 42299657 94M2S
#> 6 29223687 42299657 94M2S

H3122_results$last_EML4
#> EML4_seq
#> CCAGGCTGGAGTGCAGTGGT GGAGTGCAGTGGTGTGATTT TCAGGCTGGAGTGCAGTGGT
#>                  201                    1                    1

H3122_results$first_ALK
#> ALK_seq
#> CAGAATTTTAGCTTTGCAAT      CGGATTTTTAGCTTT CGGATTTTTAGCTTTTCATT
#>                    1                    1                    2
#> CGGAATTTTAGCTTTGCATT                   CT                  CTG
#>                    1                    8                    3
#>                 CTGA                CTGAT             CTGATTTT
#>                   11                   16                    5
#>            CTGATTTTT           CTGATTTTTA          CTGATTTTTAG
#>                    6                    3                    3
#> CTGATTTTTAGATTTGCATT         CTGATTTTTAGC        CTGATTTTTAGCT
#>                    1                   14                   10
#>       CTGATTTTTAGCTT      CTGATTTTTAGCTTT     CTGATTTTTAGCTTTG
#>                   10                    3                    4
#>    CTGATTTTTAGCTTTGC   CTGATTTTTAGCTTTGCA  CTGATTTTTAGCTTTGCAT
#>                    7                    8                    1
#> CTGATTTTTAGCTTTGCATT CTGATTTTTAGCTTTGCAAT     CTGATTTTTAGCTTTT
#>                   71                    1                    1
#> CTGATTTTTAGCTTTTCATA          CTGATTTTTAT     CTGATTTTTATCTTTG
#>                    1                    1                    2
#> CTGATTTTTATCTTTGCATT CTGATTTTTATCTTTTGATT CTGTGTTTTAGATTTGCATT
#>                    2                    1                    1
#> CTGTTTTTTATCTTTGCAAT                CTGAA CTTATTTTTATCTTTGCATT
#>                    1                    1                    1
#>            TTAGCTTTG
#>                    1

H3122_results$breakpoint
#> break_pos
#> 42299750 42299757
#>      202        1

H3122_results$read_depth
#> [1] 251

HCC827_results
#> [1] "No EML4-ALK was detected"
```