Genefusiondiscover

Christoffer Trier Maansson and Emma Roger Andersen

09 aug 2022

Abstract

Circulating tumor DNA (ctDNA) containing somatic mutations can be found in blood plasma. This includes DNA fusions, such as the EML4-ALK, which can be an oncogenic driver in non-small cell lung cancer.This is an introduction to the **Genefusiondiscover** package for R, which can be used to evaluate whether EML4-ALK is present in blood plasma.

Contents

1 Introduction

This package was created in order to increase the sensitivity of EML4-ALK detection from commercially available NGS products such as 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 the 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.

2 Installation

Use **devtools** to install the most recent version of **Genefusiondiscover** from the GitHub repository.

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

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

Alternatively, install **Genefusiondiscover** published at **Bioconductor**

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("Genefusiondiscover")
library(Genefusiondiscover)
```

3 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 = "Genefusiondiscover")
HCC827_bam <-  system.file("extdata",
                            "HCC827_EML4.bam",
                            package = "Genefusiondiscover")
```

4 Functions

4.1 `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 either be "hg38" or "hg19".
Default = "hg38".
```

`mates`

```
integer, 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
#> 726 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 727 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 728 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 729 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 731 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 733 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#>         mate position cigar
#> 726 29223691 42299657 94M2S
#> 727 29223375 42299657 94M2S
#> 728 29223479 42299657 94M2S
#> 729 29223686 42299657 94M2S
#> 731 29223636 42299657 94M2S
#> 733 29223687 42299657 94M2S
```

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

4.2 `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, a `table` of identified EML4 basepairs is returned, 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"
```

4.3 `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, a `table` of identified ALK basepairs is returned, 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 CGGAATTTTAGCTTTGCATT      CGGATTTTTAGCTTT
#>                    1                    1                    1
#> CGGATTTTTAGCTTTTCATT                   CT                  CTG
#>                    2                    8                    3
#>                 CTGA                CTGAA                CTGAT
#>                   11                    1                   16
#>             CTGATTTT            CTGATTTTT           CTGATTTTTA
#>                    5                    6                    3
#>          CTGATTTTTAG CTGATTTTTAGATTTGCATT         CTGATTTTTAGC
#>                    3                    1                   14
#>        CTGATTTTTAGCT       CTGATTTTTAGCTT      CTGATTTTTAGCTTT
#>                   10                   10                    3
#>     CTGATTTTTAGCTTTG    CTGATTTTTAGCTTTGC   CTGATTTTTAGCTTTGCA
#>                    4                    7                    8
#> CTGATTTTTAGCTTTGCAAT  CTGATTTTTAGCTTTGCAT CTGATTTTTAGCTTTGCATT
#>                    1                    1                   71
#>     CTGATTTTTAGCTTTT CTGATTTTTAGCTTTTCATA          CTGATTTTTAT
#>                    1                    1                    1
#>     CTGATTTTTATCTTTG CTGATTTTTATCTTTGCATT CTGATTTTTATCTTTTGATT
#>                    2                    2                    1
#> CTGTGTTTTAGATTTGCATT CTGTTTTTTATCTTTGCAAT 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"
```

4.4 `breakPosition()`

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

Examples:

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

4.5 `breakPositionDepth()`

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:

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

4.6 `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`

```
integer, 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:

```
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)
```

```
head(H3122_results$clipped_reads)
#>                                                                                              sequences
#> 726 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 727 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 728 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 729 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 731 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#> 733 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTGTTTGTTTTTTGAGATGGGGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAGTGGTCT
#>         mate position cigar
#> 726 29223691 42299657 94M2S
#> 727 29223375 42299657 94M2S
#> 728 29223479 42299657 94M2S
#> 729 29223686 42299657 94M2S
#> 731 29223636 42299657 94M2S
#> 733 29223687 42299657 94M2S
```

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

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

5 Session info

```
#> - Session info ----------------------------------------------------------------
#>  setting  value
#>  version  R version 4.2.1 (2022-06-23 ucrt)
#>  os       Windows 10 x64 (build 22000)
#>  system   x86_64, mingw32
#>  ui       RTerm
#>  language (EN)
#>  collate  C
#>  ctype    Danish_Denmark.utf8
#>  tz       Europe/Paris
#>  date     2022-08-09
#>  pandoc   2.17.1.1 @ C:/Program Files/RStudio/bin/quarto/bin/ (via rmarkdown)
#>
#> - Packages -------------------------------------------------------------------
#>  package            * version date (UTC) lib source
#>  BiocStyle          * 2.25.0  2022-04-28 [3] Bioconductor
#>  devtools           * 2.4.4   2022-07-20 [3] CRAN (R 4.2.1)
#>  dplyr              * 1.0.9   2022-04-28 [3] CRAN (R 4.2.0)
#>  Genefusiondiscover * 0.99.0  2022-08-09 [1] Bioconductor
#>  usethis            * 2.1.6   2022-05-25 [3] CRAN (R 4.2.1)
#>
#>  [1] C:/Users/Christoffer/AppData/Local/Temp/RtmpA3xrG7/Rinst71a02e6c245d
#>  [2] C:/Users/Christoffer/AppData/Local/Temp/RtmpySQZqD/temp_libpath407053e73508
#>  [3] C:/Users/Christoffer/AppData/Local/R/win-library/4.2
#>  [4] C:/Program Files/R/R-4.2.1/library
#>
#> -------------------------------------------------------------------------------
```