

Genefusiondiscover

<https://github.com/CTrierMaansson/Genefusiondiscover>

Christoffer Trier Maansson

Emma Roger Andersen

26 jul 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	2
2	Installation	2
3	Package data	2
4	Functions	3
4.1	EML4_ALK_detection()	3
4.2	EML4_sequence()	4
4.3	ALK_sequence()	4
4.4	break_position()	5
4.5	break_position_depth()	6
4.6	EML4_ALK_analysis()	6

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 **Genefusiondiscover**.

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

install_github("CTrierMaansson/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
#> 1 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTTGTTTTGTTTTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#> 2 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTTGTTTTGTTTTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#> 3 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTTGTTTTGTTTTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#> 4 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTTGTTTTGTTTTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#> 5 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTTGTTTTGTTTTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#> 6 TTGCTTCTTTCACTTAGTTTTTTTTGTTTTGTTTTGTTTTGTTTTGTTTTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#>      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"
```

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 GGAGTGCAGTGGTGTGATT 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      CGGATTTTCTAGCTTT CGGATTTTCTAGCTTTTCATT
#>          1          1          2
#> CGGAATTTTAGCTTTGCATT      CT          CTG
#>          1          8          3
#>          CTGA          CTGAT          CTGATTTT
#>          11          16          5
#>          CTGATTTT      CTGATTTTCTA      CTGATTTTCTAG
#>          6          3          3
#> CTGATTTTCTAGATTTGCATT      CTGATTTTCTAGC      CTGATTTTCTAGCT
#>          1          14          10
#>          CTGATTTTCTAGCTT      CTGATTTTCTAGCTTT      CTGATTTTCTAGCTTTG
#>          10          3          4
#>          CTGATTTTCTAGCTTTGC      CTGATTTTCTAGCTTTGCA      CTGATTTTCTAGCTTTGCAT
#>          7          8          1
#> CTGATTTTCTAGCTTTGCATT CTGATTTTCTAGCTTTGCAAT      CTGATTTTCTAGCTTTT
#>          71          1          1
#> CTGATTTTCTAGCTTTTCATA      CTGATTTTCTAT      CTGATTTTCTATCTTTG
#>          1          1          2
#> CTGATTTTCTATCTTTGCATT CTGATTTTCTATCTTTTGATT CTGTGTTTCTAGATTTGCATT
#>          2          1          1
#> CTGTGTTTCTATCTTTGCAAT      CTGAA      CTTATTTTCTATCTTTGCATT
#>          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 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, 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:

```

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"

```

4.5 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"
```

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)
```

```
#>
#> 1 TTGCTTCTTTCACTTAGTTTTTTTGTGTTTTGTTTGTGTTTTGTTTGTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#> 2 TTGCTTCTTTCACTTAGTTTTTTTGTGTTTTGTTTGTGTTTTGTTTGTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#> 3 TTGCTTCTTTCACTTAGTTTTTTTGTGTTTTGTTTGTGTTTTGTTTGTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#> 4 TTGCTTCTTTCACTTAGTTTTTTTGTGTTTTGTTTGTGTTTTGTTTGTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#> 5 TTGCTTCTTTCACTTAGTTTTTTTGTGTTTTGTTTGTGTTTTGTTTGTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#> 6 TTGCTTCTTTCACTTAGTTTTTTTGTGTTTTGTTTGTGTTTTGTTTGTGAGATGGGGTTTCACTCTTGTTGCCAGGCTGGAGTGCAGTGGTCT
#>
#>      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 GGAGTGCAGTGGTGTGATT TCAGGCTGGAGTGCAGTGGT
#>
#>      201      1      1
```

```
H3122_results$first_ALK
```

```
#> ALK_seq
#> CAGAATTTTAGCTTTGCAAT      CGGATTTTAGCTTT CGGATTTTAGCTTTTCATT
#>      1      1      2
#> CGGAATTTTAGCTTTGCATT      CT      CTG
#>      1      8      3
#>      CTGA      CTGAT      CTGATTTT
#>      11      16      5
#>      CTGATTTT      CTGATTTTTA      CTGATTTTtag
#>      6      3      3
#> CTGATTTTtagATTGCAAT      CTGATTTTtagC      CTGATTTTtagCT
#>      1      14      10
#>      CTGATTTTtagCTT      CTGATTTTtagCTTT      CTGATTTTtagCTTTG
#>      10      3      4
#>      CTGATTTTtagCTTGC      CTGATTTTtagCTTTGCA      CTGATTTTtagCTTTGCAT
#>      7      8      1
#> CTGATTTTtagCTTTGCAAT CTGATTTTtagCTTTGCAAT      CTGATTTTtagCTTTT
#>      71      1      1
```

```

#> CTGATTTTTAGCTTTTCATA          CTGATTTTTAT          CTGATTTTTATCTTTG
#>              1              1              2
#> CTGATTTTTATCTTTGCATT CTGATTTTTATCTTTGATT CTGTGTTTATAGATTGCATT
#>              2              1              1
#> CTGTTTTTATCTTTGCAAT          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"

session_info(
  pkgs = "loaded",
  include_base = FALSE,
  dependencies = NA,
  to_file = FALSE
)
#> - 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      Danish_Denmark.utf8
#> ctype        Danish_Denmark.utf8
#> tz           Europe/Paris
#> date         2022-07-26
#> pandoc       2.17.1.1 @ C:/Program Files/RStudio/bin/quarto/bin/ (via rmarkdown)
#>
#> - Packages -----
#> package      * version   date (UTC) lib source
#> assertthat   0.2.1     2019-03-21 [1] CRAN (R 4.2.0)
#> bamsignals    1.29.0    2022-04-28 [1] Bioconductor
#> BiocGenerics  0.43.0    2022-04-28 [1] Bioconductor
#> BiocParallel  1.31.10   2022-07-07 [1] Bioconductor
#> Biostrings    2.65.1    2022-06-09 [1] Bioconductor
#> bitops        1.0-7     2021-04-24 [1] CRAN (R 4.2.0)
#> cachem        1.0.6     2021-08-19 [1] CRAN (R 4.2.0)
#> callr         3.7.1     2022-07-13 [1] CRAN (R 4.2.1)
#> cli           3.3.0     2022-04-25 [1] CRAN (R 4.2.0)
#> codetools     0.2-18    2020-11-04 [2] CRAN (R 4.2.1)
#> crayon        1.5.1     2022-03-26 [1] CRAN (R 4.2.0)
#> curl          4.3.2     2021-06-23 [1] CRAN (R 4.2.0)
#> DBI           1.1.3     2022-06-18 [1] CRAN (R 4.2.1)

```



```

#> devtools          * 2.4.4    2022-07-20 [1] CRAN (R 4.2.1)
#> digest            0.6.29    2021-12-01 [1] CRAN (R 4.2.0)
#> dplyr             * 1.0.9    2022-04-28 [1] CRAN (R 4.2.0)
#> ellipsis          0.3.2     2021-04-29 [1] CRAN (R 4.2.0)
#> evaluate          0.15      2022-02-18 [1] CRAN (R 4.2.0)
#> fansi             1.0.3     2022-03-24 [1] CRAN (R 4.2.0)
#> fastmap           1.1.0     2021-01-25 [1] CRAN (R 4.2.0)
#> fs                1.5.2     2021-12-08 [1] CRAN (R 4.2.0)
#> Genefusiondiscover * 1.2      2022-07-26 [1] Github (CTrierMaansson/Genefusiondiscover@b31f350)
#> generics          0.1.3     2022-07-05 [1] CRAN (R 4.2.1)
#> GenomeInfoDb      1.33.3    2022-05-11 [1] Bioconductor
#> GenomeInfoDbData  1.2.8     2022-05-12 [1] Bioconductor
#> GenomicRanges     1.49.0    2022-04-28 [1] Bioconductor
#> glue              1.6.2     2022-02-24 [1] CRAN (R 4.2.0)
#> htmltools         0.5.3     2022-07-18 [1] CRAN (R 4.2.1)
#> htmlwidgets       1.5.4     2021-09-08 [1] CRAN (R 4.2.0)
#> httpuv            1.6.5     2022-01-05 [1] CRAN (R 4.2.0)
#> IRanges           2.31.0    2022-04-28 [1] Bioconductor
#> knitr             1.39      2022-04-26 [1] CRAN (R 4.2.0)
#> later             1.3.0     2021-08-18 [1] CRAN (R 4.2.0)
#> lifecycle         1.0.1     2021-09-24 [1] CRAN (R 4.2.0)
#> magrittr          2.0.3     2022-03-30 [1] CRAN (R 4.2.0)
#> memoise           2.0.1     2021-11-26 [1] CRAN (R 4.2.0)
#> mime              0.12      2021-09-28 [1] CRAN (R 4.2.0)
#> miniUI            0.1.1.1   2018-05-18 [1] CRAN (R 4.2.1)
#> pillar            1.8.0     2022-07-18 [1] CRAN (R 4.2.1)
#> pkgbuild          1.3.1     2021-12-20 [1] CRAN (R 4.2.1)
#> pkgconfig         2.0.3     2019-09-22 [1] CRAN (R 4.2.0)
#> pkgload           1.3.0     2022-06-27 [1] CRAN (R 4.2.1)
#> prettyunits       1.1.1     2020-01-24 [1] CRAN (R 4.2.0)
#> processx          3.7.0     2022-07-07 [1] CRAN (R 4.2.1)
#> profvis           0.3.7     2020-11-02 [1] CRAN (R 4.2.1)
#> promises          1.2.0.1   2021-02-11 [1] CRAN (R 4.2.0)
#> ps                1.7.1     2022-06-18 [1] CRAN (R 4.2.1)
#> purrr             0.3.4     2020-04-17 [1] CRAN (R 4.2.0)
#> R6                2.5.1     2021-08-19 [1] CRAN (R 4.2.0)
#> Rcpp              1.0.9     2022-07-08 [1] CRAN (R 4.2.1)
#> RCurl             1.98-1.7   2022-06-09 [1] CRAN (R 4.2.0)
#> remotes           2.4.2     2021-11-30 [1] CRAN (R 4.2.0)
#> rlang             1.0.4     2022-07-12 [1] CRAN (R 4.2.1)
#> rmarkdown         2.14      2022-04-25 [1] CRAN (R 4.2.0)
#> Rsamtools         2.13.3    2022-05-25 [1] Bioconductor
#> rstudioapi        0.13      2020-11-12 [1] CRAN (R 4.2.0)
#> S4Vectors          0.35.1    2022-06-08 [1] Bioconductor
#> sessioninfo       1.2.2     2021-12-06 [1] CRAN (R 4.2.1)
#> shiny             1.7.2     2022-07-19 [1] CRAN (R 4.2.1)
#> stringi           1.7.8     2022-07-11 [1] CRAN (R 4.2.1)
#> stringr           1.4.0     2019-02-10 [1] CRAN (R 4.2.0)
#> tibble            3.1.8     2022-07-22 [1] CRAN (R 4.2.1)
#> tidyselect        1.1.2     2022-02-21 [1] CRAN (R 4.2.0)
#> urlchecker        1.0.1     2021-11-30 [1] CRAN (R 4.2.1)
#> usethis           * 2.1.6     2022-05-25 [1] CRAN (R 4.2.1)
#> utf8              1.2.2     2021-07-24 [1] CRAN (R 4.2.0)
#> vctrs             0.4.1     2022-04-13 [1] CRAN (R 4.2.0)

```

```
#> withr          2.5.0    2022-03-03 [1] CRAN (R 4.2.0)
#> xfun           0.31     2022-05-10 [1] CRAN (R 4.2.0)
#> xtable         1.8-4    2019-04-21 [1] CRAN (R 4.2.0)
#> XVector        0.37.0   2022-04-28 [1] Bioconductor
#> yaml           2.3.5    2022-02-21 [1] CRAN (R 4.2.0)
#> zlibbioc       1.43.0   2022-04-28 [1] Bioconductor
#>
#> [1] C:/Users/Christoffer/AppData/Local/R/win-library/4.2
#> [2] C:/Program Files/R/R-4.2.1/library
#>
#> -----
```