

An Introduction to *FilterFFPE*

Lanying Wei

Modified: 20 August, 2020. Compiled: May 19, 2021

Contents

1	Introduction	1
2	Input	1
3	Artifact Chimeric Read Filtration	1
A	Session info	3

1 Introduction

The next-generation sequencing (NGS) reads from formalin-fixed paraffin-embedded (FFPE) samples contain numerous artifact chimeric reads, which can lead to a large number of false positive structural variation (SV) calls. The *FilterFFPE* package finds and filters these artifact chimeric reads from BAM files to improve SV calling performance in FFPE samples.

2 Input

The required input is an indexed BAM file of the FFPE sample, the PCR or optical duplicates should be marked or removed from the BAM file. If you plan to filter reads with mapping quality, or to only keep reads in targeted region, please do that after using *FilterFFPE*, as these steps may remove some of the alignments that is useful to the *FilterFFPE*'s filtering algorithm. Example of such a BAM file is stored in the 'extdata' directory of *FilterFFPE* package).

3 Artifact Chimeric Read Filtration

The filtration includes two steps: 1) Find artifact chimeric reads from BAM file . 2) Remove these artifact chimeric reads in the filtered BAM file. `findArtifactChimericReads` can be used to find artifact chimeric reads; read names of PCR or optical duplicates of all chimeric reads are also found and written in a txt file for filtration. `filterBamByReadNames` can be used for further filtration, it generates a filtered and indexed BAM file. `FFPEReadFilter` combines these two functions.

```
> library(FilterFFPE)
> # Find artifact chimeric reads
> file <- system.file("extdata", "example.bam", package = "FilterFFPE")
```

```
> outFolder <- tempdir()
> FFPEReadsFile <- paste0(outFolder, "/example.FFPEReads.txt")
> dupChimFile <- paste0(outFolder, "/example.dupChim.txt")
> artifactReads <- findArtifactChimericReads(file = file, threads = 2,
+                                           FFPEReadsFile = FFPEReadsFile,
+                                           dupChimFile = dupChimFile)
> head(artifactReads)

[1] "SRR1523265.24545253" "SRR1523265.31420529"
[3] "SRR1523265.38291385" "SRR1523265.49620943"
[5] "SRR1523265.5056364"  "SRR1523265.52887917"

>
```

```
> # Filter artifact chimeric reads and PCR or optical duplicates of chimeric reads
> dupChim <- readLines(dupChimFile)
> readsToFilter <- c(artifactReads, dupChim)
> destination <- paste0(outFolder, "/example.FilterFFPE.bam")
> filterBamByReadNames(file = file, readsToFilter = readsToFilter,
+                      destination = destination, overwrite=TRUE)

[1] "/tmp/RtmppnovUv/example.FilterFFPE.bam"

>
```

```
> # Perform finding and filtering with one function
> file <- system.file("extdata", "example.bam", package = "FilterFFPE")
> outFolder <- tempdir()
> FFPEReadsFile <- paste0(outFolder, "/example.FFPEReads.txt")
> dupChimFile <- paste0(outFolder, "/example.dupChim.txt")
> destination <- paste0(outFolder, "/example.FilterFFPE.bam")
> FFPEReadFilter(file = file, threads=2, destination = destination,
+               overwrite=TRUE, FFPEReadsFile = FFPEReadsFile,
+               dupChimFile = dupChimFile)

[1] "/tmp/RtmppnovUv/example.FilterFFPE.bam"

>
```

The generated BAM file can be loaded with `scanBam` function from *Rsamtools* package for further interrogation.

```
> # load Bam file with scanBAM
> newBam <- Rsamtools::scanBam(destination)
> head(newBam[[1]]$seq)

DNAStringSet object of length 6:
      width seq
[1]   90 CAGCTGCTCAACCACTCCTCTCT...CCCTGGCCCTCCCAGCCCACGAT
[2]   90 CAGCTGCTCAACCACTCCTCTCT...CCCTGGCCCTCCCAGCCCACGAT
[3]   90 CAGCTGCTCAACCACTCCTCTCT...CCCTGGCCCTCCCAGCCCACGAT
[4]   90 CAGCTGCTCAACCACTCCTCTCT...CCCTGGCCCTCCCAGCCCACGAT
[5]   90 ACCCACTCCCTGGCCCTCCCAGC...CCTGAACCCCCAGCCTGTGGTTC
[6]   90 CCCCCACTCCCTGGCCCTCCCAGC...CCTGAACCCCCAGCCTGTGGTTC
```

A Session info

```
> packageDescription("FilterFFPE")

Package: FilterFFPE
Type: Package
Title: FFPE Artificial Chimeric Read Filter for NGS
      data
Version: 1.3.0
Authors@R: person("Lanying", "Wei",
  email="lanying.wei@uni-muenster.de", role =
    c("aut", "cre"), comment = c(ORCID =
      "0000-0002-4281-8017"))
Description: This package finds and filters
  artificial chimeric reads specifically
  generated in next-generation sequencing (NGS)
  process of formalin-fixed paraffin-embedded
  (FFPE) tissues. These artificial chimeric reads
  can lead to a large number of false positive
  structural variation (SV) calls. The required
  input is an indexed BAM file of a FFPE sample.
License: LGPL-3
Encoding: UTF-8
Imports: foreach, doParallel, GenomicRanges, IRanges,
  Rsamtools, parallel, S4Vectors
Suggests: BiocStyle
biocViews: StructuralVariation, Sequencing,
  Alignment, QualityControl, Preprocessing
git_url:
  https://git.bioconductor.org/packages/FilterFFPE
git_branch: master
git_last_commit: 7608296
git_last_commit_date: 2021-05-19
Date/Publication: 2021-05-19
Author: Lanying Wei [aut, cre]
  (<https://orcid.org/0000-0002-4281-8017>)
Maintainer: Lanying Wei <lanying.wei@uni-muenster.de>
Built: R 4.1.0; ; 2021-05-19 21:16:41 UTC; unix

-- File: /tmp/RtmpgFFnPI/Rinst133cba2e873732/FilterFFPE/Meta/package.rds

> sessionInfo()

R version 4.1.0 beta (2021-05-03 r80259)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS

Matrix products: default
BLAS: /home/biocbuild/bbs-3.14-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.14-bioc/R/lib/libRlapack.so

locale:
```

```
[1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
[3] LC_TIME=en_GB             LC_COLLATE=C
[5] LC_MONETARY=en_US.UTF-8   LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8      LC_NAME=C
[9] LC_ADDRESS=C              LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

attached base packages:

```
[1] stats      graphics  grDevices  utils      datasets
[6] methods    base
```

other attached packages:

```
[1] FilterFFPE_1.3.0
```

loaded via a namespace (and not attached):

```
[1] rstudioapi_0.13      knitr_1.33
[3] XVector_0.33.0       GenomicRanges_1.45.0
[5] BiocGenerics_0.39.0   zlibbioc_1.39.0
[7] IRanges_2.27.0        BiocParallel_1.27.0
[9] doParallel_1.0.16     rlang_0.4.11
[11] foreach_1.5.1         GenomeInfoDb_1.29.0
[13] tools_4.1.0           parallel_4.1.0
[15] xfun_0.23             iterators_1.0.13
[17] htmltools_0.5.1.1     yaml_2.2.1
[19] digest_0.6.27         crayon_1.4.1
[21] GenomeInfoDbData_1.2.6 BiocManager_1.30.15
[23] codetools_0.2-18      S4Vectors_0.31.0
[25] bitops_1.0-7          RCurl_1.98-1.3
[27] evaluate_0.14         rmarkdown_2.8
[29] compiler_4.1.0        Rsamtools_2.9.0
[31] Biostrings_2.61.0     stats4_4.1.0
[33] BiocStyle_2.21.0
```