In_silico_digestion_and_GFF_processing

Violeta de Anca & Fábio Pertille

2022-12-08

Load all the packages

These packages are required to run this script.

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

```
## Warning: package 'BiocManager' was built under R version 4.2.2
```

```
## Bioconductor version '3.15' is out-of-date; the current release version '3.16'
##   is available with R version '4.2'; see https://bioconductor.org/install
```

```
BiocManager::install("BSgenome.Mmusculus.UCSC.mm39")
```

```
## Bioconductor version 3.15 (BiocManager 1.30.19), R 4.2.0 (2022-04-22 ucrt)
```

```
## Warning: package(s) not installed when version(s) same as or greater than current; use
##   `force = TRUE` to re-install: 'BSgenome.Mmusculus.UCSC.mm39'
```

```
## Installation paths not writeable, unable to update packages
##   path: C:/Program Files/R/R-4.2.0/library
##   packages:
##     boot, cluster, foreign, Matrix, mgcv, nlme, nnet, rpart, survival
```

```
## Old packages: 'cli', 'digest', 'edgeR', 'htmltools', 'limma', 'MASS', 'plyr',
##   'rbibutils', 'roxygen2', 'segmented', 'testthat', 'xfun'
```

```
BiocManager::install("rtracklayer", force = TRUE)
```

```
## Bioconductor version 3.15 (BiocManager 1.30.19), R 4.2.0 (2022-04-22 ucrt)
```

```
## Installing package(s) 'rtracklayer'
```

```
## package 'rtracklayer' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\viode560\AppData\Local\Temp\Rtmpw1D56X\downloaded_packages
```

```
## Installation paths not writeable, unable to update packages
##   path: C:/Program Files/R/R-4.2.0/library
##   packages:
##     boot, cluster, foreign, Matrix, mgcv, nlme, nnet, rpart, survival

## Old packages: 'cli', 'digest', 'edgeR', 'htmltools', 'limma', 'MASS', 'plyr',
##   'rbibutils', 'roxygen2', 'segmented', 'testthat', 'xfun'
```

```
BiocManager::install("GenomicRanges", force = TRUE)
```

```
## Bioconductor version 3.15 (BiocManager 1.30.19), R 4.2.0 (2022-04-22 ucrt)

## Installing package(s) 'GenomicRanges'

## package 'GenomicRanges' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\viode560\AppData\Local\Temp\Rtmpw1D56X\downloaded_packages

## Installation paths not writeable, unable to update packages
##   path: C:/Program Files/R/R-4.2.0/library
##   packages:
##     boot, cluster, foreign, Matrix, mgcv, nlme, nnet, rpart, survival

## Old packages: 'cli', 'digest', 'edgeR', 'htmltools', 'limma', 'MASS', 'plyr',
##   'rbibutils', 'roxygen2', 'segmented', 'testthat', 'xfun'
```

```
BiocManager::install("genomation")
```

```
## Bioconductor version 3.15 (BiocManager 1.30.19), R 4.2.0 (2022-04-22 ucrt)

## Warning: package(s) not installed when version(s) same as or greater than current; use
##   'force = TRUE' to re-install: 'genomation'

## Installation paths not writeable, unable to update packages
##   path: C:/Program Files/R/R-4.2.0/library
##   packages:
##     boot, cluster, foreign, Matrix, mgcv, nlme, nnet, rpart, survival
## Old packages: 'cli', 'digest', 'edgeR', 'htmltools', 'limma', 'MASS', 'plyr',
##   'rbibutils', 'roxygen2', 'segmented', 'testthat', 'xfun'
```

```
BiocManager::install("Rsubread")
```

```
## Bioconductor version 3.15 (BiocManager 1.30.19), R 4.2.0 (2022-04-22 ucrt)

## Warning: package(s) not installed when version(s) same as or greater than current; use
##   'force = TRUE' to re-install: 'Rsubread'
```

```
## Installation paths not writeable, unable to update packages
##   path: C:/Program Files/R/R-4.2.0/library
##   packages:
##     boot, cluster, foreign, Matrix, mgcv, nlme, nnet, rpart, survival
## Old packages: 'cli', 'digest', 'edgeR', 'htmltools', 'limma', 'MASS', 'plyr',
##   'rbibutils', 'roxygen2', 'segmented', 'testthat', 'xfun'
```

```
library(Biostrings)
```

```
## Warning: package 'Biostrings' was built under R version 4.2.1

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

## Warning: package 'IRanges' was built under R version 4.2.1

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows
```

```
## Loading required package: XVector

## Loading required package: GenomeInfoDb

## Warning: package 'GenomeInfoDb' was built under R version 4.2.1

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':
##
##     strsplit
```

```
library(Rsubread)
```

```
## Warning: package 'Rsubread' was built under R version 4.2.1
```

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 4.2.2
```

```
library(genomation)
```

```
## Loading required package: grid

##
## Attaching package: 'grid'

## The following object is masked from 'package:Biostrings':
##
##     pattern

## Warning: replacing previous import 'Biostrings::pattern' by 'grid::pattern' when
## loading 'genomation'
```

```
library(GenomicRanges)
library(rtracklayer)
library(BSgenome.Mmusculus.UCSC.mm39)
```

```
## Loading required package: BSgenome
```

```
library(tibble)
```

```
## Warning: package 'tibble' was built under R version 4.2.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union
```

```
## The following objects are masked from 'package:Biostrings':
##
##     collapse, intersect, setdiff, setequal, union
```

```
## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect
```

```
## The following object is masked from 'package:XVector':
##
##     slice
```

```
## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union
```

```
## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union
```

```
## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.2.2
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.2.1
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:S4Vectors':
##
##     expand
```

Constructing PstI digestion in silico in Mus musculus mm39

In this chunk of the code we are creating gaps for the motif of the PstI, then extracting the fragments between the gaps, getting the coordinates where the gaps start and end. As these coordinates do not consider the cut of the PstI, coordinates need to be corrected. Finally we show the first 6 lines of the dataframe.

```
mdf=data.frame();
for (i in seq_along(Mmusculus)){
  m<-matchPattern(c("CTGCAG"), Mmusculus[[i]])
  starts<-start(gaps(m))
  ends<-end(gaps(m))
  temp_df<-data.frame(start=starts-5,end=ends+1,chr=seqnames(Mmusculus)[i])
  temp_df$start<-replace(temp_df$start, temp_df$start <= 0, 1)
  mdf<-rbind(mdf,temp_df)
}
head(mdf)
```

```
##     start     end  chr
## 1       1 3054610 chr1
## 2 3054611 3056435 chr1
## 3 3056436 3065279 chr1
## 4 3065280 3066729 chr1
## 5 3066730 3069367 chr1
## 6 3069368 3073783 chr1
```

Formatting the file to adapt between USCS and NCBI annotation formats

In USCS format it is not used chr for the annotation, FeatureCounts needs to have the same annotation in both GFF file and in the BAM files so it can recognize the windows.

```
mdf["chr"] = str_remove(mdf$chr, "chr")
mdf$start=gsub('^0$', '1', mdf$start)
head(mdf)
```

```
##     start     end chr
## 1       1 3054610   1
## 2 3054611 3056435   1
## 3 3056436 3065279   1
## 4 3065280 3066729   1
## 5 3066730 3069367   1
## 6 3069368 3073783   1
```

Now export directly the bed file as gff3 format

```
export(mdf, "in_silico_windows_mm39.gff", format = "gtf")
```