Extraction of Barcodes and gene fragments

Tomas Bjorklund

Thu Oct 29 09:57:50 2020

This workflow extracts barcodes and the gene fragments synthesized with the CustomArray using bbmap2. The fragments are then suitable for alignment to reference sequences using blastn.

```r
suppressPackageStartupMessages(library(knitr))
```

Sequencing files

```r
dataDir <- config$Value[1]
in.name.P5 <- file.path(dataDir, config$Value[2])
in.name.P7 <- file.path(dataDir, config$Value[3])
name.out <- config$Value[4]
paired.alignment <- as.logical(config$Value[5])
```

Analysis parameters

```r
knitr::kable(config, format = "latex", booktabs = T) %>% kable_styling(latex_options = "striped")
```

Parameter	Value
dataDir	seqFiles
in.name.P5	DNA_pscAAVlib_R1.fastq.gz
in.name.P7	DNA_pscAAVlib_R2.fastq.gz
name.out	AAVlibrary_complete
paired.alignment	TRUE
run.subset	FALSE
max.cores	32
subset.count	250000

```r
run.subset <- as.logical(config$Value[6])
max.cores <- as.integer(config$Value[7])
subset.count <- as.integer(config$Value[8])

strt <- Sys.time()
```

Selection of real amplicons

```r
# This section searches the sequencing file and only select the files with
# valid amplicons
out.name.P5 <- tempfile(pattern = "P5_", tmpdir = tempdir(), fileext = ".fastq.gz")
out.name.P7 <- tempfile(pattern = "P7_", tmpdir = tempdir(), fileext = ".fastq.gz")
command.args <- paste("overwrite=true k=15 rcomp=f skipr2=t qhdist=0 maskmiddle=f",
    " hammingdistance=2 findbestmatch=f ordered=t threads=", detectCores(),
    " in=", in.name.P5, " in2=", in.name.P7, " outm=", out.name.P5, " outm2=",
    out.name.P7, " fliteral=", "GTATGTTGTTCTGGAGCGGGAGGGTGCTATTTTGCCTAGCGATAA",
    sep = "")

sys.out <- system2(path.expand("~/bbmap/bbduk2.sh"), args = command.args, stdout = TRUE,
    stderr = TRUE)

sys.out <- as.data.frame(sys.out)

colnames(sys.out) <- c("bbduk2 Identification of real amplicons")
invisible(sys.out[" "] <- " ")
lengthOut <- (nrow(sys.out))
knitr::kable(sys.out[3:lengthOut, ], format = "latex", booktabs = T) %>% kable_styling(latex_options = "str
```

	bbduk2 Identification of real amplicons
3	
4	BBDuk2 version 37.02
5	Set ORDERED to true
6	Set threads to 48
7	k=15
8	hamming distance=2
9	kfiltering using 1 literal.
10	
11	Initial:
12	Memory: max=50358m, free=49045m, used=1313m
13	
14	Added 30721 kmers; time: 0.092 seconds.
15	Memory: max=50358m, free=47206m, used=3152m
16	
17	Input is being processed as paired
18	Started output streams: 0.263 seconds.
19	Processing time: 126.836 seconds.
20	
21	Input: 23191088 reads 3490215687 bases.
22	Contaminants: 23095890 reads (99.59%) 3475908371 bases (99.59%)
23	Total Removed: 23095890 reads (99.59%) 3475908371 bases (99.59%)
24	Result: 95198 reads (0.41%) 14307316 bases (0.41%)
25	
26	Time: 127.210 seconds.
27	Reads Processed: 23191k 182.31k reads/sec
28	Bases Processed: 3490m 27.44m bases/sec

```
in.name.P5 <- out.name.P5
in.name.P7 <- out.name.P7
```

Extraction of a subset

```
if (run.subset) {
    suppressWarnings(sampler <- FastqSampler(gsub("([\\])", "", in.name.P5),
        subset.count, readerBlockSize = 1e+09, ordered = TRUE))
    set.seed(123)
    tmp.P5 <- yield(sampler)
    in.name.P5 <- tempfile(pattern = "P5_", tmpdir = tempdir(), fileext = ".fastq.gz")
    writeFastq(tmp.P5, in.name.P5, compress = TRUE)
    rm(tmp.P5)
    suppressWarnings(sampler <- FastqSampler(gsub("([\\])", "", in.name.P7),
        subset.count, readerBlockSize = 1e+09, ordered = TRUE))
    set.seed(123)
    tmp.P7 <- yield(sampler)
    in.name.P7 <- tempfile(pattern = "P7_", tmpdir = tempdir(), fileext = ".fastq.gz")
    writeFastq(tmp.P7, in.name.P7, compress = TRUE)
    rm(tmp.P7)
}

output.Reads <- as.integer(system(paste("gunzip -c ", shQuote(gsub("([\\])",
    "", in.name.P5)), " | echo $((`wc -l`/4)) 2>&1", sep = ""), intern = TRUE,
    ignore.stdout = FALSE))  #Stores the read count utilized
print(paste("Utilized sequences:", output.Reads))
```

```
[1] "Utilized sequences: 11547945"
```

Extraction of barcodes

```
out.name.P5 <- tempfile(pattern = "BC_", tmpdir = tempdir(), fileext = ".fastq.gz")

sys.out <- system(paste("~/bbmap/bbduk2.sh overwrite=true k=18 mink=18 hammingdistance=2 findbestmatch=t ",
    "rcomp=f findbestmatch=f qhdist=1 minavgquality=0 maxns=0 minlength=18 ",
    "maxlength=22 threads=", detectCores(), " in=", shQuote(in.name.P5), " out=",
    out.name.P5, " lliteral=", "GGCCTAGCGGCCGCTTTACTT", " rliteral=", "ATAACTTCGTATAATGTATGC",
    " 2>&1", sep = ""), intern = TRUE, ignore.stdout = FALSE)
sys.out <- as.data.frame(sys.out)

in.name.P5 <- out.name.P5


colnames(sys.out) <- c("bbduk2 Extraction of barcodes")
invisible(sys.out[" "] <- " ")
lengthOut <- (nrow(sys.out))
knitr::kable(sys.out[3:lengthOut, ], format = "latex", booktabs = T) %>% kable_styling(latex_options = "str
```

```
rm(sys.out)

reads.BC <- readFastq(in.name.P5)
sread(reads.BC)
```

```
  A DNAStringSet instance of length 11374106
           width seq
       [1]    20 GTTTCAGTACAGGCGGATTT
       [2]    20 ATGGCAACCAGGGAATGCGC
       [3]    20 GTGTCATGGATTATACGTAT
       [4]    20 ACTGGTATGTTGGCACACTC
       [5]    20 GTACATGTATATATGTCTAC
       ...   ... ...
[11374102]    20 GCGCGTTGATGGGTTGGCTC
[11374103]    20 CAGGCCTGAATGGCGGGCGG
[11374104]    20 GTTTGCTAGTTCCCGGGTAC
[11374105]    20 CAGGGCGGGTGTGCAGGCGG
[11374106]    20 GTACGTTTCCGTCCATATTG
```

	bbduk2 Extraction of barcodes
3	
4	BBDuk2 version 37.02
5	Set threads to 48
6	k=18
7	maskMiddle=true
8	hamming distance=2
9	right-ktrimming using 1 literal.
10	left-ktrimming using 1 literal.
11	
12	Initial:
13	Memory: max=50341m, free=47975m, used=2366m
14	
15	Added 5104 kmers; time: 0.051 seconds.
16	Memory: max=50341m, free=46136m, used=4205m
17	
18	Added 5104 kmers; time: 0.020 seconds.
19	Memory: max=50341m, free=46136m, used=4205m
20	
21	Input is being processed as unpaired
22	Started output streams: 0.099 seconds.
23	Processing time: 259.354 seconds.
24	
25	Input: 11547945 reads 1738273591 bases.
26	KTrimmed: 23030031 reads (199.43%) 1505373909 bases (86.60%)
27	Low quality discards: 6 reads (0.00%) 122 bases (0.00%)
28	Total Removed: 173839 reads (1.51%) 1511015923 bases (86.93%)
29	Result: 11374106 reads (98.49%) 227257668 bases (13.07%)
30	
31	Time: 259.545 seconds.
32	Reads Processed: 11547k 44.49k reads/sec
33	Bases Processed: 1738m 6.70m bases/sec

```r
(unique.BCs <- unique(sread(reads.BC)))
```

```
  A DNAStringSet instance of length 3934570
          width seq
      [1]    20 GTTTCAGTACAGGCGGATTT
      [2]    20 ATGGCAACCAGGGAATGCGC
      [3]    20 GTGTCATGGATTATACGTAT
      [4]    20 ACTGGTATGTTGGCACACTC
      [5]    20 GTACATGTATATATGTCTAC
      ...   ... ...
[3934566]    20 GCTCCCGGGAAGCTTCCCGT
[3934567]    20 AAATACTGGCTGATAACCTG
[3934568]    20 GCATCCTTATTTCATGCTTT
[3934569]    20 GCGCGCTGATGTGTTCGCGG
[3934570]    20 GTTTGCTAGTTCCCGGGTAC
```

```r
output.BCs <- length(unique.BCs)
print(paste("Utilized barcodes:", output.BCs))
```

```
[1] "Utilized barcodes: 3934570"
```

```
barcodeTable <- data.table(ID = as.character(ShortRead::id(reads.BC)), BC = as.character(sread(reads.BC)))
```

Extraction of fragments

```
out.name.P7 <- tempfile(pattern = "P7_", tmpdir = tempdir(), fileext = ".fastq.gz")
command.args <- paste("overwrite=true k=18 mink=18 rcomp=f qhdist=1 maskmiddle=t",
    " hammingdistance=2 findbestmatch=t minlength=38 maxlength=78 ordered=t ",
    "threads=", detectCores(), " in=", in.name.P7, " out=", out.name.P7, " lliteral=",
    "AGCAACCTCCAGAGAGGCAACG", " rliteral=", "CAGACAAGCAGCTACCGCAGAT", sep = "")

sys.out <- system2(path.expand("~/bbmap/bbduk2.sh"), args = command.args, stdout = TRUE,
    stderr = TRUE)  #

sys.out <- as.data.frame(sys.out)

colnames(sys.out) <- c("bbduk2 extraction of fragments")
invisible(sys.out[" "] <- " ")
lengthOut <- (nrow(sys.out))
knitr::kable(sys.out[3:lengthOut, ], format = "latex", booktabs = T) %>% kable_styling(latex_options = "str
```

```
in.name.P7 <- out.name.P7

out.name.P5 <- tempfile(pattern = "P5_", tmpdir = tempdir(), fileext = ".fastq.gz")
out.name.P7 <- tempfile(pattern = "P7_", tmpdir = tempdir(), fileext = ".fastq.gz")
out.name.P5_singlet <- tempfile(pattern = "P5_singlet_", tmpdir = tempdir(),
    fileext = ".fastq.gz")
out.name.P7_singlet <- tempfile(pattern = "P7_singlet_", tmpdir = tempdir(),
    fileext = ".fastq.gz")

command.args <- paste("makepairs -c 'gzip' -f ", in.name.P5, " -r ", in.name.P7,
    " -fp ", out.name.P5, " -rp ", out.name.P7, " -fs ", out.name.P5_singlet,
    " -rs ", out.name.P7_singlet, " --stats 2>&1", sep = "")
sys.out <- system2("/usr/local/bin/pairfq", args = command.args, stdout = TRUE,
    stderr = TRUE)
sys.out <- as.data.frame(sys.out)

colnames(sys.out) <- c("pairfq pair matching")
invisible(sys.out[" "] <- " ")
lengthOut <- (nrow(sys.out))
knitr::kable(sys.out[1:lengthOut, ], format = "latex", booktabs = T) %>% kable_styling(latex_options = "str
```

```
rm(sys.out)


system(paste("mv ", out.name.P5, " ./data/barcodes_", name.out, ".fastq.gz",
    sep = ""))
system(paste("mv ", out.name.P7, " ./data/fragments_", name.out, ".fastq.gz",
    sep = ""))

unlink(paste(tempdir(), "/*", sep = ""), recursive = FALSE, force = FALSE)  #Cleanup of temp files

print("Total execution time:")
```

```
[1] "Total execution time:"
```

	bbduk2 extraction of fragments
3	
4	BBDuk2 version 37.02
5	Set ORDERED to true
6	Set threads to 48
7	k=18
8	maskMiddle=true
9	hamming distance=2
10	right-ktrimming using 1 literal.
11	left-ktrimming using 1 literal.
12	
13	Initial:
14	Memory: max=48918m, free=46874m, used=2044m
15	
16	Added 6380 kmers; time: 0.053 seconds.
17	Memory: max=48918m, free=45087m, used=3831m
18	
19	Added 6380 kmers; time: 0.020 seconds.
20	Memory: max=48918m, free=45087m, used=3831m
21	
22	Input is being processed as unpaired
23	Started output streams: 0.095 seconds.
24	Processing time: 266.683 seconds.
25	
26	Input: 11547945 reads 1737634780 bases.
27	KTrimmed: 22321617 reads (193.30%) 1078844314 bases (62.09%)
28	Total Removed: 698793 reads (6.05%) 1146258928 bases (65.97%)
29	Result: 10849152 reads (93.95%) 591375852 bases (34.03%)
30	
31	Time: 266.871 seconds.
32	Reads Processed: 11547k 43.27k reads/sec
33	Bases Processed: 1737m 6.51m bases/sec

pairfq pair matching
========= pairfq version : 0.17.0 (completion time: Thu Oct 29 10:19:17 UTC 2020)
Total forward reads (/tmp/Rtmpd0f3N4/BC__2bd64e93472.fastq.gz) : 11374106
Total reverse reads (/tmp/Rtmpd0f3N4/P7__2bd45d1c469.fastq.gz) : 10849152
Total forward paired reads (/tmp/Rtmpd0f3N4/P5__2bd759c717a.fastq.gz) : 10698072
Total reverse paired reads (/tmp/Rtmpd0f3N4/P7__2bdac335f9.fastq.gz) : 10698072
Total forward unpaired reads (/tmp/Rtmpd0f3N4/P5__singlet__2bd7391d943.fastq.gz) : 676034
Total reverse unpaired reads (/tmp/Rtmpd0f3N4/P7__singlet__2bd73472390.fastq.gz) : 151080
Total paired reads : 21396144
Total unpaired reads : 827114

```
print(Sys.time() - strt)
```

```
Time difference of 21.30765 mins
```

```
devtools::session_info()
```

```
Session info -----------------------------------------------------------
```

```
 setting  value
 version  R version 3.4.2 (2017-09-28)
 system   x86_64, linux-gnu
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 tz       UTC
 date     2020-10-29

Packages -----------------------------------------------------------------

 package              * version  date       source
 acepack                1.4.1    2016-10-29 CRAN (R 3.4.2)
 backports              1.1.1    2017-09-25 CRAN (R 3.4.2)
 base                 * 3.4.2    2017-10-06 local
 base64enc              0.1-3    2015-07-28 CRAN (R 3.4.2)
 Biobase              * 2.36.2   2017-11-29 Bioconductor
 BiocGenerics         * 0.22.1   2017-11-29 Bioconductor
 BiocParallel         * 1.10.1   2017-11-29 Bioconductor
 Biostrings           * 2.44.2   2017-11-29 Bioconductor
 bitops                 1.0-6    2013-08-17 CRAN (R 3.4.2)
 checkmate              1.8.4    2017-09-25 CRAN (R 3.4.2)
 cluster                2.0.6    2017-03-16 CRAN (R 3.4.2)
 codetools              0.2-15   2016-10-05 CRAN (R 3.4.2)
 colorspace             1.3-2    2016-12-14 CRAN (R 3.4.2)
 compiler               3.4.2    2017-10-06 local
 data.table           * 1.10.4-2 2017-10-12 url
 datasets             * 3.4.2    2017-10-06 local
 DelayedArray         * 0.2.7    2017-11-29 Bioconductor
 devtools             * 1.13.3   2017-08-02 CRAN (R 3.4.2)
 digest                 0.6.12   2017-01-27 CRAN (R 3.4.2)
 doParallel           * 1.0.11   2017-09-28 CRAN (R 3.4.2)
 evaluate               0.10.1   2017-06-24 CRAN (R 3.4.2)
 foreach              * 1.4.3    2015-10-13 CRAN (R 3.4.2)
 foreign                0.8-69   2017-06-21 CRAN (R 3.4.2)
 formatR                1.5      2017-04-25 CRAN (R 3.4.2)
 Formula              * 1.2-2    2017-07-10 CRAN (R 3.4.2)
 GenomeInfoDb         * 1.12.3   2017-11-29 Bioconductor
 GenomeInfoDbData       0.99.0   2017-11-29 Bioconductor
 GenomicAlignments    * 1.12.2   2017-11-29 Bioconductor
 GenomicRanges        * 1.28.6   2017-11-29 Bioconductor
 ggplot2              * 2.2.1    2016-12-30 CRAN (R 3.4.2)
 graphics             * 3.4.2    2017-10-06 local
 grDevices            * 3.4.2    2017-10-06 local
 grid                   3.4.2    2017-10-06 local
 gridExtra              2.3      2017-09-09 CRAN (R 3.4.2)
 gtable                 0.2.0    2016-02-26 CRAN (R 3.4.2)
 Hmisc                * 4.0-3    2017-05-02 CRAN (R 3.4.2)
 hms                    0.3      2016-11-22 CRAN (R 3.4.2)
 htmlTable              1.9      2017-01-26 CRAN (R 3.4.2)
 htmltools              0.3.6    2017-04-28 CRAN (R 3.4.2)
 htmlwidgets            0.9      2017-07-10 CRAN (R 3.4.2)
 httr                   1.3.1    2017-08-20 CRAN (R 3.4.2)
 hwriter                1.3.2    2014-09-10 CRAN (R 3.4.2)
 IRanges              * 2.10.5   2017-11-29 Bioconductor
 iterators            * 1.0.8    2015-10-13 CRAN (R 3.4.2)
 kableExtra           * 0.5.2    2017-09-15 url
 knitr                * 1.17     2017-08-10 CRAN (R 3.4.2)
```

```
lattice              * 0.20-35  2017-03-25 CRAN (R 3.4.2)
latticeExtra           0.6-28   2016-02-09 CRAN (R 3.4.2)
lazyeval               0.2.0    2016-06-12 CRAN (R 3.4.2)
magrittr               1.5      2014-11-22 CRAN (R 3.4.2)
Matrix                 1.2-11   2017-08-21 url
matrixStats          * 0.52.2   2017-04-14 CRAN (R 3.4.2)
memoise                1.1.0    2017-04-21 CRAN (R 3.4.2)
methods              * 3.4.2    2017-10-06 local
munsell                0.4.3    2016-02-13 CRAN (R 3.4.2)
nnet                   7.3-12   2016-02-02 CRAN (R 3.4.2)
parallel             * 3.4.2    2017-10-06 local
plyr                   1.8.4    2016-06-08 CRAN (R 3.4.2)
R6                     2.2.2    2017-06-17 CRAN (R 3.4.2)
RColorBrewer           1.1-2    2014-12-07 CRAN (R 3.4.2)
Rcpp                   0.12.13  2017-09-28 url
RCurl                  1.95-4.8 2016-03-01 CRAN (R 3.4.2)
readr                  1.1.1    2017-05-16 CRAN (R 3.4.2)
rlang                  0.1.2    2017-08-09 CRAN (R 3.4.2)
rmarkdown              1.6      2017-06-15 url
rpart                  4.1-11   2017-04-21 CRAN (R 3.4.2)
rprojroot              1.2      2017-01-16 CRAN (R 3.4.2)
Rsamtools            * 1.28.0   2017-11-29 Bioconductor
rvest                  0.3.2    2016-06-17 CRAN (R 3.4.2)
S4Vectors            * 0.14.7   2017-11-29 Bioconductor
scales                 0.5.0    2017-08-24 CRAN (R 3.4.2)
ShortRead            * 1.34.2   2017-11-29 Bioconductor
splines                3.4.2    2017-10-06 local
stats                * 3.4.2    2017-10-06 local
stats4               * 3.4.2    2017-10-06 local
stringi                1.1.5    2017-04-07 url
stringr                1.2.0    2017-02-18 CRAN (R 3.4.2)
SummarizedExperiment * 1.6.5    2017-11-29 Bioconductor
survival             * 2.41-3   2017-04-04 CRAN (R 3.4.2)
tibble                 1.3.4    2017-08-22 CRAN (R 3.4.2)
tools                  3.4.2    2017-10-06 local
utils                * 3.4.2    2017-10-06 local
withr                  2.0.0    2017-07-28 url
xml2                   1.1.1    2017-01-24 CRAN (R 3.4.2)
XVector              * 0.16.0   2017-11-29 Bioconductor
yaml                   2.1.14   2016-11-12 CRAN (R 3.4.2)
zlibbioc               1.22.0   2017-11-29 Bioconductor
```