Barcoded extraction and reduction from RNA samples

Tomas Bjorklund

Mon Nov 2 07:13:26 2020

This workflow identifies correct amplicons from in vivo & in vitro samples and extracts the barcode. Barcodes are then reduced using the starcode algorithm.

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

Analyze tissue RNA

```
strt <- Sys.time()
load("data/multipleContfragmentsComplete.rda")
load("data/alignedLibraries.rda")
load("data/LUTdna.rda")

load.list <- read.table("input/loadlist.txt", header = FALSE, skip = 0, sep = "\t",
    stringsAsFactors = FALSE, fill = TRUE)

dataDir <- "seqFiles"
colnames(load.list) <- c("Name", "BaseName", "GroupName")

log.table <- data.table(Name = "Name", Reads = NA, Purity = NA, BCs = NA, SCdroppedBC = NA,
    allBCs = NA, scBCs = NA)

analyzeTissue <- function(indexNr) {
    # indexNr <- 1

    name <- unlist(strsplit(load.list$BaseName[indexNr], "/"))
    name <- name[!is.na(name)]
    if (length(name) == 2) {
        in.files <- list.files(paste(gsub("([\\])", "", dataDir), name[1], sep = "/"),
            pattern = paste(name[2], "*", sep = ""), full.names = TRUE)
    } else {
        in.files <- list.files(gsub("([\\])", "", dataDir), pattern = paste(name[1],
            "*", sep = ""), full.names = TRUE)
    }
    in.files.P5 <- in.files[grep("R1", in.files)]
    in.files.P7 <- in.files[grep("R2", in.files)]

    in.name.P5 <- tempfile(pattern = "P5_", tmpdir = tempdir(), fileext = ".fastq.gz")
    in.name.P7 <- tempfile(pattern = "P7_", tmpdir = tempdir(), fileext = ".fastq.gz")
    system(paste("cat '", paste(as.character(in.files.P5), collapse = "' '"),
        "' > ", in.name.P5, " 2>&1", sep = ""), intern = TRUE, ignore.stdout = FALSE)

    system(paste("cat '", paste(as.character(in.files.P7), collapse = "' '"),
        "' > ", in.name.P7, " 2>&1", sep = ""), intern = TRUE, ignore.stdout = FALSE)

    log.table$Name <- load.list$Name[indexNr]
    name.out <- log.table$Name

    # Selection of real 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=10 rcomp=f skipr1=t qhdist=0 maskmiddle=t ",
    "hammingdistance=1 findbestmatch=t ordered=t threads=", detectCores(),
    " in=", in.name.P5, " in2=", in.name.P7, " outm=", out.name.P5, " outm2=",
    out.name.P7, " fliteral=", "CGCCACAACATCGAGGACGGCAGCGTG", sep = "")

sys.out <- system2(path.expand("~/bbmap/bbduk2.sh"), args = command.args,
    stdout = TRUE, stderr = TRUE)
log.table$Purity <- strsplit(sys.out[grep("Contaminants", sys.out)], split = "\t")[[1]][2]

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

log.table$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


# Extraction of barcodes ============================

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

sys.out <- system(paste("~/bbmap/bbduk2.sh overwrite=true k=12 mink=12 hammingdistance=2 ",
    "findbestmatch=t trd=t rcomp=f skipr2=t findbestmatch=f qhdist=0 ",
    "minavgquality=0 ordered=t maxns=0 minlength=18 maxlength=22 threads=",
    detectCores(), " in=", shQuote(in.name.P5), " out=", out.name.BC, " lliteral=",
    "GGCCTAGCGGCCGCTTTACTT", " rliteral=", "ATAACTTCGTATA", " 2>&1", sep = ""),
    intern = TRUE, ignore.stdout = FALSE)

log.table$BCs <- strsplit(sys.out[grep("Result:", sys.out)], split = "\t")[[1]][2]

reads.BC <- readFastq(out.name.BC)
barcodeTable <- data.table(ID = as.character(ShortRead::id(reads.BC)), BC = as.character(sread(reads.BC)
    key = "BC")

# Starcode based barcode reduction ============================

out.name.BC.star <- tempfile(pattern = "BCsc_", tmpdir = tempdir(), fileext = ".txt")

system(paste("gunzip -c ", out.name.BC, " | starcode -t ", detectCores(),
    " --print-clusters -d", 1, " -r5 -q -o ", out.name.BC.star, " 2>&1",
    sep = ""), intern = TRUE, ignore.stdout = FALSE)

table.BC.sc <- data.table(read.table(out.name.BC.star, header = FALSE, row.names = 1,
    skip = 0, sep = "\t", stringsAsFactors = FALSE, fill = FALSE), keep.rownames = TRUE,
    key = "rn")
table.BC.sc[, `:=`(V2, NULL)]

table.BC.sc <- table.BC.sc[, strsplit(as.character(V3), ",", fixed = TRUE),
    by = rn]

log.table$SCdroppedBC <- length(unique(sread(reads.BC))) - length(unique(table.BC.sc$V1) %in%
    unique(sread(reads.BC)))


setnames(table.BC.sc, c("V1", "rn"), c("BC", "scBC"))
```

```
    # Replacing barcodes with Starcode reduced versions
    # ============================


    setkey(table.BC.sc, BC)

    barcodeTable <- barcodeTable[table.BC.sc, nomatch = 0]

    setnames(barcodeTable, c("BC", "scBC"), c("oldBC", "BC"))

    setkey(barcodeTable, BC)

    log.table$allBCs <- length(unique(barcodeTable$oldBC))
    log.table$scBCs <- length(unique(barcodeTable$BC))

    invisible(barcodeTable[, `:=`(oldBC, NULL)])
    setkey(output.Table, "BC")

    BCcount <- data.table(as.data.frame(rev(sort(table(barcodeTable$BC))), row.names = "Var1"),
        keep.rownames = TRUE)
    # In R versions below 3.3 remove, row.names = 'Var1' to make this compatible
    setnames(BCcount, colnames(BCcount), c("BC", "RNAcount"))
    setkey(BCcount, "BC")
    foundFrags <- output.Table[BCcount, nomatch = 0]
    setkey(foundFrags, "LUTnr")
    setkey(LUT.dna, "LUTnr")
    foundFrags <- foundFrags[LUT.dna, nomatch = 0]
    setnames(foundFrags, "Sequence", "fragment")
    foundFrags[, `:=`(c("Names", "i.Structure"), NULL)]

    matchRange <- function(idxFrag) {
        # idxFrag <- 23
        matchRanges <- which(mcols(allFragments.ranges)$Sequence == foundFrags$fragment[idxFrag])
        return(cbind(matchRanges, idxFrag))
    }
    match.ranges.list <- mclapply(1:nrow(foundFrags), matchRange, mc.preschedule = TRUE,
        mc.cores = detectCores()/2)
    match.ranges <- do.call(rbind, match.ranges.list)
    foundFragments.ranges <- allFragments.ranges[match.ranges[, 1]]
    if (ncol(match.ranges) >= 2) {
        foundFrags <- foundFrags[match.ranges[, "idxFrag"], ]
        foundFrags[, `:=`(c("Reads", "fragment", "Structure", "LUTnr"), NULL)]
        mcols(foundFragments.ranges) <- c(mcols(foundFragments.ranges), foundFrags)
        o = order(-mcols(foundFragments.ranges)$RNAcount)
        foundFragments.ranges <- foundFragments.ranges[o]
        saveRDS(foundFragments.ranges, file = paste("output/", "found.", name.out,
            ".rds", sep = ""), compress = TRUE)
    }
    return(log.table)
}
```

Analysis summary

```
all.logs <- lapply(1:nrow(load.list), analyzeTissue)
all.logs <- rbindlist(all.logs, use.names = FALSE)
knitr::kable(all.logs, format = "latex", longtable = T, booktabs = T) %>% kable_styling(latex_options = c("s
    "scale_down", "repeat_header")) %>% landscape()
```

```
Warning in styling_latex_scale_down(out, table_info): Longtable cannot be
resized.
```

Name	Reads	Purity	BCs	SCdroppedBC	allBCs	scBCs
DNA_pscAAVlib_Prep2	41210335	82420670 reads (99.65%)	23986369 reads (58.20%)	60046	4940426	3938187
DNA_AAVlib_DNAse_3cpc	7964874	15929748 reads (99.66%)	2985701 reads (37.49%)	154	341674	223158
DNA_AAVlib_DNAse_30cpc	17557643	35115286 reads (99.69%)	8827549 reads (50.28%)	2464	1157571	795330
mRNA_30cpc_SN_RatNr7	1983137	3966274 reads (99.65%)	1614014 reads (81.39%)	41	24691	9962
mRNA_30cpc_Ctx_RatNr7	1994467	3988934 reads (99.60%)	1771085 reads (88.80%)	13	16108	7547
mRNA_30cpc_Th_RatNr7	2867972	5735944 reads (99.65%)	1382945 reads (48.22%)	12	30889	11632
mRNA_30cpc_Str_RatNr7	1596468	3192936 reads (99.62%)	991144 reads (62.08%)	15	60701	21186
mRNA_30cpc_SN_RatNr1	1611759	3223518 reads (99.67%)	1361550 reads (84.48%)	47	11471	6081
mRNA_30cpc_Ctx_RatNr1	1541657	3083314 reads (99.67%)	1203038 reads (78.04%)	9	11542	7747
mRNA_30cpc_Th_RatNr1	2359538	4719076 reads (99.67%)	1026288 reads (43.50%)	1	17976	8587
mRNA_30cpc_Str_RatNr1	1505088	3010176 reads (99.60%)	619506 reads (41.16%)	6	37192	13397
mRNA_30cpc_SN_RatNr8	2054931	4109862 reads (99.70%)	1215848 reads (59.17%)	80	22685	11414
mRNA_30cpc_Ctx_RatNr8	2102816	4205632 reads (99.69%)	1081123 reads (51.41%)	18	13376	7115
mRNA_30cpc_Th_RatNr8	2105768	4211536 reads (99.69%)	856547 reads (40.68%)	9	29605	12444
mRNA_30cpc_Str_RatNr8	1623686	3247372 reads (99.64%)	612950 reads (37.75%)	9	45404	20175
mRNA_3cpc_SN_RatNr15	1436268	2872536 reads (99.64%)	1435201 reads (99.93%)	17	4226	1538
mRNA_3cpc_Ctx_RatNr15	1260242	2520484 reads (99.65%)	1062278 reads (84.29%)	12	6434	3211
mRNA_3cpc_Th_RatNr15	1105966	2211932 reads (99.65%)	944716 reads (85.42%)	3	5642	2818
mRNA_3cpc_Str_RatNr15	948187	1896374 reads (99.65%)	684628 reads (72.20%)	2	22301	6617
mRNA_3cpc_SN_RatNr21	1115267	2230534 reads (99.75%)	1112638 reads (99.76%)	3	4489	2021
mRNA_3cpc_Ctx_RatNr21	1201263	2402526 reads (99.73%)	788753 reads (65.66%)	2	3920	1926
mRNA_3cpc_Th_RatNr21	1234915	2469830 reads (99.73%)	1063068 reads (86.08%)	0	7414	3780
mRNA_3cpc_Str_RatNr21	1151549	2303098 reads (99.71%)	845867 reads (73.45%)	3	29847	8603
mRNA_3cpc_Ctx_RatNr19	1743952	3487904 reads (99.61%)	1738647 reads (99.70%)	15	6568	2765
mRNA_3cpc_Th_RatNr19	1738722	3477444 reads (99.68%)	1174178 reads (67.53%)	21	10753	6535
mRNA_3cpc_Str_RatNr19	1645506	3291012 reads (99.67%)	1196277 reads (72.70%)	5	34666	12990
mRNA_3cpc_Th_RatNr20	788088	1576176 reads (99.61%)	682035 reads (86.54%)	3	5098	2470
mRNA_3cpc_Str_RatNr20	983263	1966526 reads (99.59%)	827836 reads (84.19%)	3	23426	8334
mRNA_30cpc_Organoid_MD114	25382593	50765186 reads (99.68%)	18874513 reads (74.36%)	162	140815	54807
mRNA_3000cpc_Organoid_MD101	36282473	72564946 reads (99.57%)	16907669 reads (46.60%)	516	325836	127004
mRNA_3cpc_HEK293Nr2	1588150	3176300 reads (99.63%)	1016131 reads (63.98%)	11	7051	3030
mRNA_30cpc_HEK293Nr3	1920390	3840780 reads (99.65%)	864987 reads (45.04%)	13	20920	7730
mRNA_3cpc_pNeuronNr6	1143395	2286790 reads (99.56%)	552817 reads (48.35%)	4	6313	2553
mRNA_30cpc_pNeuronNr7	1652464	3304928 reads (99.66%)	942060 reads (57.01%)	13	24858	8032
mRNA_30cpc_4wks_Ctx_RatNr2	1955085	3910170 reads (99.25%)	591146 reads (30.24%)	3	5607	2561
mRNA_30cpc_4wks_SN_RatNr2	1921383	3842766 reads (99.39%)	1210727 reads (63.01%)	10	18161	8012
mRNA_30cpc_4wks_Str_RatNr2	2101122	4202244 reads (100.00%)	1563614 reads (74.42%)	24	84742	59375
mRNA_30cpc_4wks_Th_RatNr2	2177483	4354966 reads (95.47%)	752026 reads (34.54%)	0	16068	5602

(continued)

Name	Reads	Purity	BCs	SCdroppedBC	allBCs	scBCs
mRNA_3cpc_4wks_Ctx_RatNr13	2070763	4141526 reads (99.61%)	183573 reads (8.86%)	1	1403	674
mRNA_3cpc_4wks_SN_RatNr13	1809233	3618466 reads (99.47%)	105370 reads (5.82%)	2	1696	1249
mRNA_3cpc_4wks_Str_RatNr13	1693037	3386074 reads (99.32%)	1096589 reads (64.77%)	2	17281	6400
mRNA_3cpc_4wks_Th_RatNr13	1954529	3909058 reads (99.61%)	1079929 reads (55.25%)	27	12266	5208

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

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

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

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

```
Time difference of 1.302718 hours
```

```
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     2020-11-02
```

```
Packages -----------------------------------------------------------------
```

```
 package               * version  date       source
 acepack                 1.4.1    2016-10-29 CRAN (R 3.4.2)
 AnnotationDbi         * 1.38.2   2017-11-29 Bioconductor
 AnnotationFilter        1.0.0    2017-11-29 Bioconductor
 AnnotationHub           2.8.3    2017-11-29 Bioconductor
 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)
 beanplot              * 1.2      2014-09-19 CRAN (R 3.4.2)
 Biobase               * 2.36.2   2017-11-29 Bioconductor
 BiocGenerics          * 0.22.1   2017-11-29 Bioconductor
 BiocInstaller           1.26.1   2017-10-10 Bioconductor
 BiocParallel          * 1.10.1   2017-11-29 Bioconductor
 biomaRt                 2.32.1   2017-11-29 Bioconductor
 Biostrings            * 2.44.2   2017-11-29 Bioconductor
 biovizBase            * 1.24.0   2017-11-29 Bioconductor
 bit                     1.1-12   2014-04-09 CRAN (R 3.4.2)
 bit64                   0.9-7    2017-05-08 CRAN (R 3.4.2)
 bitops                  1.0-6    2013-08-17 CRAN (R 3.4.2)
 blob                    1.1.0    2017-06-17 CRAN (R 3.4.2)
 BSgenome              * 1.44.2   2017-11-29 Bioconductor
 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
 curl                    2.8.1    2017-07-21 CRAN (R 3.4.2)
 data.table            * 1.10.4-2 2017-10-12 url
 datasets              * 3.4.2    2017-10-06 local
 DBI                     0.7      2017-06-18 CRAN (R 3.4.2)
 DelayedArray          * 0.2.7    2017-11-29 Bioconductor
 devtools              * 1.13.3   2017-08-02 CRAN (R 3.4.2)
 dichromat               2.0-0    2013-01-24 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)
```

```
ensembldb                2.0.4    2017-11-29 Bioconductor
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
GenomicFeatures        * 1.28.5   2017-11-29 Bioconductor
GenomicRanges          * 1.28.6   2017-11-29 Bioconductor
GGally                   1.3.2    2017-08-02 CRAN (R 3.4.2)
ggbio                  * 1.24.1   2017-11-29 Bioconductor
ggplot2                * 2.2.1    2016-12-30 CRAN (R 3.4.2)
graph                    1.54.0   2017-11-29 Bioconductor
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)
Gviz                   * 1.20.0   2017-11-29 Bioconductor
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)
httpuv                   1.3.5    2017-07-04 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)
interactiveDisplayBase   1.14.0   2017-11-29 Bioconductor
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
mime                     0.5      2016-07-07 CRAN (R 3.4.2)
munsell                  0.4.3    2016-02-13 CRAN (R 3.4.2)
nnet                     7.3-12   2016-02-02 CRAN (R 3.4.2)
OrganismDbi              1.18.1   2017-11-29 Bioconductor
parallel               * 3.4.2    2017-10-06 local
plyr                   * 1.8.4    2016-06-08 CRAN (R 3.4.2)
ProtGenerics             1.8.0    2017-11-29 Bioconductor
R6                       2.2.2    2017-06-17 CRAN (R 3.4.2)
RBGL                     1.52.0   2017-11-29 Bioconductor
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)
reshape                  0.8.7    2017-08-06 CRAN (R 3.4.2)
reshape2                 1.4.2    2016-10-22 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
RSQLite                  2.0      2017-06-19 CRAN (R 3.4.2)
rtracklayer            * 1.36.6   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)
shiny                    1.0.5    2017-08-23 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
VariantAnnotation        1.22.3   2017-11-29 Bioconductor
withr                    2.0.0    2017-07-28 url
XML                      3.98-1.9 2017-06-19 CRAN (R 3.4.2)
xml2                     1.1.1    2017-01-24 CRAN (R 3.4.2)
xtable                   1.8-2    2016-02-05 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
```