# Reverse mapping of CustumArray oligos to original proteins

*Tomas Bjorklund*

*Thu Oct 29 10:19:21 2020*

This workflow aligns the short oligos from the CustomArray order to the full reference sequences using Bowtie2. This enables the mapping to all genes sharing the same sequence.

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

## Load sequences

```
LUT.dna <- read.table("data/SortedFragments_all.txt", header = TRUE, skip = 0,
    sep = "\t", stringsAsFactors = FALSE, fill = TRUE)
LUT.dna <- data.table(LUT.dna)
```

## Remove constitutive backbone sequences

```
invisible(LUT.dna[, `:=`(Sequence, gsub("aacctccagagaggcaacg", "", Sequence))])
invisible(LUT.dna[, `:=`(Sequence, gsub("cagacaagcagctaccgca", "", Sequence))])
invisible(LUT.dna[, `:=`(Sequence, toupper(Sequence))])
setkey(LUT.dna, "Sequence")
LUT.dna <- unique(LUT.dna)
LUT.dna$Names <- LUT.dna$Sequence
LUT.dna$LUTnr <- make.names(seq(nrow(LUT.dna)), unique = TRUE)
```

## Split sequences based on linker and length

```
LUT.14aaG4S <- LUT.dna[substr(LUT.dna$Sequence, 1, 14) == "GAGGCGGAGGAAGT"]
LUT.remaining <- LUT.dna[!(substr(LUT.dna$Sequence, 1, 14) == "GAGGCGGAGGAAGT")]
LUT.14aaA5 <- LUT.remaining[substr(LUT.remaining$Sequence, 1, 14) == "CTGCTGCAGCAGCC"]
LUT.remaining <- LUT.remaining[!(substr(LUT.remaining$Sequence, 1, 14) == "CTGCTGCAGCAGCC")]
LUT.22aa <- LUT.remaining[nchar(LUT.remaining$Sequence) == 70L & substr(LUT.remaining$Sequence,
    1, 2) == "CT"]
LUT.remaining <- LUT.remaining[!(nchar(LUT.remaining$Sequence) == 70L & substr(LUT.remaining$Sequence,
    1, 2) == "CT")]
LUT.14aa <- LUT.remaining[nchar(LUT.remaining$Sequence) == 46L & substr(LUT.remaining$Sequence,
    1, 2) == "CT"]
rm(LUT.remaining)
LUT.dna[LUT.dna$Sequence %in% LUT.14aaG4S$Sequence, "Structure"] <- "14aaG4S"
LUT.dna[LUT.dna$Sequence %in% LUT.14aaA5$Sequence, "Structure"] <- "14aaA5"
LUT.dna[LUT.dna$Sequence %in% LUT.22aa$Sequence, "Structure"] <- "22aa"
LUT.dna[LUT.dna$Sequence %in% LUT.14aa$Sequence, "Structure"] <- "14aa"

save(LUT.dna, file = "data/LUTdna.rda")
```

## Trim sequences

```r
LUT.14aa$Sequence <- substr(LUT.14aa$Sequence, 3, 44)
LUT.14aaG4S$Sequence <- substr(LUT.14aaG4S$Sequence, 15, 56)
LUT.14aaA5$Sequence <- substr(LUT.14aaA5$Sequence, 15, 56)
LUT.22aa$Sequence <- substr(LUT.22aa$Sequence, 3, 68)
```

## Save fasta files for Bowtie alignments

```r
LUT.14aa.fa <- tempfile(pattern = "LUT_14aa_", tmpdir = tempdir(), fileext = "fa")
LUT.14aa.seq = ShortRead(DNAStringSet(LUT.14aa$Sequence), BStringSet(LUT.14aa$LUTnr))
writeFasta(LUT.14aa.seq, LUT.14aa.fa)

LUT.14aaG4S.fa <- tempfile(pattern = "LUT_14aaG4s_", tmpdir = tempdir(), fileext = "fa")
LUT.14aaG4S.seq = ShortRead(DNAStringSet(LUT.14aaG4S$Sequence), BStringSet(LUT.14aaG4S$LUTnr))
writeFasta(LUT.14aaG4S.seq, LUT.14aaG4S.fa)

LUT.14aaA5.fa <- tempfile(pattern = "LUT_14aaA5_", tmpdir = tempdir(), fileext = "fa")
LUT.14aaA5.seq = ShortRead(DNAStringSet(LUT.14aaA5$Sequence), BStringSet(LUT.14aaA5$LUTnr))
writeFasta(LUT.14aaA5.seq, LUT.14aaA5.fa)

LUT.22aa.fa <- tempfile(pattern = "LUT_14aaA5_", tmpdir = tempdir(), fileext = "fa")
LUT.22aa.seq = ShortRead(DNAStringSet(LUT.22aa$Sequence), BStringSet(LUT.22aa$LUTnr))
writeFasta(LUT.22aa.seq, LUT.22aa.fa)
```

## Build Bowtie index

```r
seqs.original <- readFasta("input/DNA-lib_RetrogradeTransport.fasta")

seqs.AA <- Biostrings::translate(sread(seqs.original), genetic.code = GENETIC_CODE,
    if.fuzzy.codon = "error")

source("functions/AAtoDNA.R")
seqs.optimized = ShortRead(DNAStringSet(sapply(seqs.AA, function(x) AAtoDNA(x,
    species = "hsa"))), BStringSet(gsub("([ ])", "_", ShortRead::id(seqs.original))))

bowtie.fasta <- tempfile(pattern = "bowtie_", tmpdir = tempdir(), fileext = ".fa")

writeFasta(seqs.optimized, bowtie.fasta)

bowtie.idx <- tempfile(pattern = "IDX_bowtie_", tmpdir = tempdir(), fileext = "")

sys.out <- system(paste("bowtie2-build", bowtie.fasta, bowtie.idx, "2>&1", sep = " "),
    intern = TRUE, ignore.stdout = FALSE)
```

## Align fragments to reference

Align 14aa sequences

```r
name.bowtie <- tempfile(pattern = "bowtie_", tmpdir = tempdir(), fileext = "")
```

```
sys.out <- system(paste("bowtie2 --non-deterministic --threads ", detectCores(),
    " --very-sensitive -f -a", " -x ", bowtie.idx, " -U ", LUT.14aa.fa, " -S ",
    name.bowtie, ".sam 2>&1", sep = ""), intern = TRUE, ignore.stdout = FALSE)


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

colnames(sys.out) <- c("Bowtie 2 alignment to library")
invisible(sys.out[" "] <- " ")
lengthOut <- (nrow(sys.out))
knitr::kable(sys.out[1:lengthOut, ], format = "latex", booktabs = T) %>% kable_styling(latex_options = "str
```

| Bowtie 2 alignment to library |
|---|
| 44705 reads; of these: |
| 44705 (100.00%) were unpaired; of these: |
| 0 (0.00%) aligned 0 times |
| 27420 (61.34%) aligned exactly 1 time |
| 17285 (38.66%) aligned >1 times |
| 100.00% overall alignment rate |

```
command.args <- paste("view -@ ", detectCores(), " -Sb ", name.bowtie, ".sam > ",
    name.bowtie, ".bam", sep = "")
system2("/usr/local/bin/samtools", args = command.args, stdout = TRUE, stderr = TRUE)
```

```
character(0)
```

```
command.args <- paste("sort -@ ", detectCores(), " ", name.bowtie, ".bam -o ",
    name.bowtie, "_sort.bam", sep = "")
system2("/usr/local/bin/samtools", args = command.args, stdout = TRUE, stderr = TRUE)
```

```
character(0)
```

```
frag14aa.ranges <- readGAlignments(paste(name.bowtie, "_sort.bam", sep = ""),
    use.names = TRUE)
length(names(frag14aa.ranges))
```

```
[1] 75152
```

```
length(unique(names(frag14aa.ranges)))
```

```
[1] 44705
```

```
length(unique(LUT.14aa$Sequence))
```

```
[1] 44705
```

Align 14aaG4S sequences

```
name.bowtie <- tempfile(pattern = "bowtie_", tmpdir = tempdir(), fileext = "")

sys.out <- system(paste("bowtie2 --non-deterministic --threads ", detectCores(),
    " --very-sensitive -f -a", " -x ", bowtie.idx, " -U ", LUT.14aaG4S.fa, " -S ",
    name.bowtie, ".sam 2>&1", sep = ""), intern = TRUE, ignore.stdout = FALSE)

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

colnames(sys.out) <- c("Bowtie 2 alignment to library")
invisible(sys.out[" "] <- " ")
lengthOut <- (nrow(sys.out))
knitr::kable(sys.out[1:lengthOut, ], format = "latex", booktabs = T) %>% kable_styling(latex_options = "str
```

| Bowtie 2 alignment to library |
|---|
| 15792 reads; of these: |
| 15792 (100.00%) were unpaired; of these: |
| 0 (0.00%) aligned 0 times |
| 9150 (57.94%) aligned exactly 1 time |
| 6642 (42.06%) aligned >1 times |
| 100.00% overall alignment rate |

```
command.args <- paste("view -@ ", detectCores(), " -Sb ", name.bowtie, ".sam > ",
    name.bowtie, ".bam", sep = "")
system2("/usr/local/bin/samtools", args = command.args, stdout = TRUE, stderr = TRUE)
```

```
character(0)
```

```
command.args <- paste("sort -@ ", detectCores(), " ", name.bowtie, ".bam -o ",
    name.bowtie, "_sort.bam", sep = "")
system2("/usr/local/bin/samtools", args = command.args, stdout = TRUE, stderr = TRUE)
```

```
character(0)
```

```
frag14aaG4S.ranges <- readGAlignments(paste(name.bowtie, "_sort.bam", sep = ""),
    use.names = TRUE)
length(names(frag14aaG4S.ranges))
```

```
[1] 27778
```

```
length(unique(names(frag14aaG4S.ranges)))
```

```
[1] 15792
```

```
length(unique(LUT.14aaG4S$Sequence))
```

```
[1] 15792
```

Align 14aaA5 sequences

```
name.bowtie <- tempfile(pattern = "bowtie_", tmpdir = tempdir(), fileext = "")

sys.out <- system(paste("bowtie2 --non-deterministic --threads ", detectCores(),
    " --very-sensitive -f -a", " -x ", bowtie.idx, " -U ", LUT.14aaA5.fa, " -S ",
    name.bowtie, ".sam 2>&1", sep = ""), intern = TRUE, ignore.stdout = FALSE)

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

colnames(sys.out) <- c("Bowtie 2 alignment to library")
invisible(sys.out[" "] <- " ")
lengthOut <- (nrow(sys.out))
knitr::kable(sys.out[1:lengthOut, ], format = "latex", booktabs = T) %>% kable_styling(latex_options = "stri
```

| Bowtie 2 alignment to library |
|---|
| 15792 reads; of these: |
| 15792 (100.00%) were unpaired; of these: |
| 0 (0.00%) aligned 0 times |
| 9150 (57.94%) aligned exactly 1 time |
| 6642 (42.06%) aligned >1 times |
| 100.00% overall alignment rate |

```
command.args <- paste("view -@ ", detectCores(), " -Sb ", name.bowtie, ".sam > ",
    name.bowtie, ".bam", sep = "")
system2("/usr/local/bin/samtools", args = command.args, stdout = TRUE, stderr = TRUE)
```

```
character(0)
```

```
command.args <- paste("sort -@ ", detectCores(), " ", name.bowtie, ".bam -o ",
    name.bowtie, "_sort.bam", sep = "")
system2("/usr/local/bin/samtools", args = command.args, stdout = TRUE, stderr = TRUE)
```

```
character(0)
```

```
frag14aaA5.ranges <- readGAlignments(paste(name.bowtie, "_sort.bam", sep = ""),
    use.names = TRUE)
length(names(frag14aaA5.ranges))
```

```
[1] 27778
```

```
length(unique(names(frag14aaA5.ranges)))
```

```
[1] 15792
```

```
length(unique(LUT.14aaA5$Sequence))
```

```
[1] 15792
```

Align 22aa sequences

```
name.bowtie <- tempfile(pattern = "bowtie_", tmpdir = tempdir(), fileext = "")

sys.out <- system(paste("bowtie2 --non-deterministic --threads ", detectCores(),
    " --very-sensitive -f -a", " -x ", bowtie.idx, " -U ", LUT.22aa.fa, " -S ",
    name.bowtie, ".sam 2>&1", sep = ""), intern = TRUE, ignore.stdout = FALSE)

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

colnames(sys.out) <- c("Bowtie 2 alignment to library")
invisible(sys.out[" "] <- " ")
lengthOut <- (nrow(sys.out))
knitr::kable(sys.out[1:lengthOut, ], format = "latex", booktabs = T) %>% kable_styling(latex_options = "stri
```

| Bowtie 2 alignment to library |
|---|
| 16054 reads; of these: |
| 16054 (100.00%) were unpaired; of these: |
| 0 (0.00%) aligned 0 times |
| 8730 (54.38%) aligned exactly 1 time |
| 7324 (45.62%) aligned >1 times |
| 100.00% overall alignment rate |

```
command.args <- paste("view -@ ", detectCores(), " -Sb ", name.bowtie, ".sam > ",
    name.bowtie, ".bam", sep = "")
system2("/usr/local/bin/samtools", args = command.args, stdout = TRUE, stderr = TRUE)
```

```
character(0)
```

```
command.args <- paste("sort -@ ", detectCores(), " ", name.bowtie, ".bam -o ",
    name.bowtie, "_sort.bam", sep = "")
system2("/usr/local/bin/samtools", args = command.args, stdout = TRUE, stderr = TRUE)
```

```
character(0)
```

```
frag22aa.ranges <- readGAlignments(paste(name.bowtie, "_sort.bam", sep = ""),
    use.names = TRUE)
length(names(frag22aa.ranges))
```

```
[1] 29665
```

```
length(unique(names(frag22aa.ranges)))
```

```
[1] 16054
```

```
length(unique(LUT.22aa$Sequence))
```

```
[1] 16054
```

## Merge and annotate aligned sequences

```
mcols(frag14aa.ranges)$structure <- "14aa"
mcols(frag22aa.ranges)$structure <- "22aa"
mcols(frag14aaA5.ranges)$structure <- "14aaA5"
mcols(frag14aaG4S.ranges)$structure <- "14aaG4S"
allFragments.ranges <- append(frag14aa.ranges, frag22aa.ranges)
allFragments.ranges <- append(allFragments.ranges, frag14aaA5.ranges)
allFragments.ranges <- append(allFragments.ranges, frag14aaG4S.ranges)


mcols(allFragments.ranges)$LUTnr <- names(allFragments.ranges)
setkey(LUT.dna, LUTnr)
mcols(allFragments.ranges)$Sequence <- LUT.dna[mcols(allFragments.ranges)$LUTnr]$Sequence

save(allFragments.ranges, file = "data/alignedLibraries.rda")

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)
 ade4                   1.7-8    2017-08-09 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)
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)
evaluate               0.10.1   2017-06-24 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)
GeneGA               * 1.26.0   2017-11-29 Bioconductor
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)
hash                 * 2.2.6    2013-02-21 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
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)
seqinr               * 3.4-5    2017-08-01 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
```