Normalize Library counts

Tomas Bjorklund

Mon Nov 2 08:33:36 2020

This workflow normalizes read counts between samples to compensate for variable read depth.

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

Generate load list and grouping names

```
strt <- Sys.time()
in.names.all <- list.files("output", pattern = "*.rds", full.names = TRUE)
load.list <- read.table("input/loadlist.txt", header = FALSE, skip = 0, sep = "\t",
    stringsAsFactors = FALSE, fill = TRUE)
colnames(load.list) <- c("Name", "BaseName", "GroupName")
load.list <- rbind(load.list, c("completeLibraryRanges", "", "DNA_pscAAVlib"))
load.list <- load.list[!grepl("Untreat", load.list$Name), ]

select.Cases <- c(unlist(sapply(load.list$Name, function(x) grep(x, in.names.all),
    simplify = TRUE)))

(in.names.all <- in.names.all[select.Cases])
```

```
 [1] "output/found.DNA_pscAAVlib_Prep2.rds"
 [2] "output/found.DNA_AAVlib_DNAse_3cpc.rds"
 [3] "output/found.DNA_AAVlib_DNAse_30cpc.rds"
 [4] "output/found.mRNA_30cpc_SN_RatNr7.rds"
 [5] "output/found.mRNA_30cpc_Ctx_RatNr7.rds"
 [6] "output/found.mRNA_30cpc_Th_RatNr7.rds"
 [7] "output/found.mRNA_30cpc_Str_RatNr7.rds"
 [8] "output/found.mRNA_30cpc_SN_RatNr1.rds"
 [9] "output/found.mRNA_30cpc_Ctx_RatNr1.rds"
[10] "output/found.mRNA_30cpc_Th_RatNr1.rds"
[11] "output/found.mRNA_30cpc_Str_RatNr1.rds"
[12] "output/found.mRNA_30cpc_SN_RatNr8.rds"
[13] "output/found.mRNA_30cpc_Ctx_RatNr8.rds"
[14] "output/found.mRNA_30cpc_Th_RatNr8.rds"
[15] "output/found.mRNA_30cpc_Str_RatNr8.rds"
[16] "output/found.mRNA_3cpc_SN_RatNr15.rds"
[17] "output/found.mRNA_3cpc_Ctx_RatNr15.rds"
[18] "output/found.mRNA_3cpc_Th_RatNr15.rds"
[19] "output/found.mRNA_3cpc_Str_RatNr15.rds"
[20] "output/found.mRNA_3cpc_SN_RatNr21.rds"
[21] "output/found.mRNA_3cpc_Ctx_RatNr21.rds"
[22] "output/found.mRNA_3cpc_Th_RatNr21.rds"
[23] "output/found.mRNA_3cpc_Str_RatNr21.rds"
[24] "output/found.mRNA_3cpc_Ctx_RatNr19.rds"
[25] "output/found.mRNA_3cpc_Th_RatNr19.rds"
[26] "output/found.mRNA_3cpc_Str_RatNr19.rds"
[27] "output/found.mRNA_3cpc_Th_RatNr20.rds"
[28] "output/found.mRNA_3cpc_Str_RatNr20.rds"
[29] "output/found.mRNA_30cpc_Organoid_MD114 mRNA_30cpc_Organoid_MD114_R mRNA_30cpc_Organoid_MD114.rds"
[30] "output/found.mRNA_30cpc_Organoid_MD114.rds"
```

```
[31] "output/found.mRNA_3000cpc_Organoid_MD101 mRNA_3000cpc_Organoid_MD101_R mRNA_3000cpc_Organoid_MD101.rds
[32] "output/found.mRNA_3000cpc_Organoid_MD101.rds"
[33] "output/found.mRNA_3cpc_HEK293Nr2.rds"
[34] "output/found.mRNA_30cpc_HEK293Nr3.rds"
[35] "output/found.mRNA_3cpc_pNeuronNr6.rds"
[36] "output/found.mRNA_30cpc_pNeuronNr7.rds"
[37] "output/found.mRNA_30cpc_4wks_Ctx_RatNr2.rds"
[38] "output/found.mRNA_30cpc_4wks_SN_RatNr2.rds"
[39] "output/found.mRNA_30cpc_4wks_Str_RatNr2.rds"
[40] "output/found.mRNA_30cpc_4wks_Th_RatNr2.rds"
[41] "output/found.mRNA_3cpc_4wks_Ctx_RatNr13.rds"
[42] "output/found.mRNA_3cpc_4wks_SN_RatNr13.rds"
[43] "output/found.mRNA_3cpc_4wks_Str_RatNr13.rds"
[44] "output/found.mRNA_3cpc_4wks_Th_RatNr13.rds"
[45] "output/completeLibraryRanges.rds"
```

```r
grouping <- data.frame(Sample = gsub("-", "_", gsub("found.|(output/)|(.rds)",
    "", in.names.all)), Group = load.list[match(names(select.Cases), load.list$Name),
    "GroupName"], stringsAsFactors = FALSE)
```

Load the desired alignment files and annotating group

```r
loadRDS <- function(in.name) {
    # in.name <- in.names.all[42]
    this.sample <- readRDS(in.name)
    this.name <- gsub("-", "_", gsub("found.|(output/)|(.rds)", "", in.name))
    this.group <- grouping[match(this.name, grouping$Sample), "Group"]
    mcols(this.sample) <- cbind(mcols(this.sample), data.frame(Sample = this.name,
        Group = this.group, stringsAsFactors = FALSE))

    return(this.sample)
}

all.samples <- lapply(in.names.all, loadRDS)

all.samples <- do.call(GAlignmentsList, unlist(all.samples))
all.samples <- cbind(unlist(all.samples))[[1]]

names(all.samples) <- make.names(names(all.samples), unique = TRUE)
length.Table <- data.table(seqnames = names(seqlengths(all.samples)), seqlength = seqlengths(all.samples),
    key = "seqnames")
all.samples <- data.table(as.data.frame(all.samples), key = "seqnames")
all.samples[, `:=`(c("strand", "qwidth", "cigar", "njunc", "end"), NULL)]
all.samples <- all.samples[length.Table]  #A data.table merge to match seqlengths to their respective seqnam
all.samples[, `:=`(c("Category", "Protein", "Origin", "Extra", "Number", "GeneName"),
    tstrsplit(seqnames, ",", fixed = TRUE))]
all.samples[, `:=`(c("seqnames", "Protein", "Origin", "Extra", "Number"), NULL)]
all.samples[, `:=`(GeneName, gsub("/|_", "-", GeneName))]
```

Normalizing read counts to correct for variable read depth

```r
setkey(all.samples, Group)
all.samples <- all.samples[RNAcount > 1, ]  #Filters out single count reads
```

```
readCounts <- all.samples[, list(GroupCount = sum(RNAcount)), by = "Group"]
readCounts[, `:=`(GroupCount, GroupCount/max(GroupCount))]
setkey(readCounts, Group)
all.samples <- all.samples[readCounts]  #Merge with normalizing factor
all.samples[, `:=`(RNAcount, RNAcount/GroupCount)]
setkey(all.samples, Mode)
all.samples <- all.samples["Def"]

setkey(all.samples, Group)
total.AAV.samples <- all.samples[Group != "DNA_pscAAVlib" & Group != "DNA_pscAAVlib_Prep2" &
    Group != "DNA_AAVlib_DNAse_3cpc" & Group != "DNA_AAVlib_DNAse_30cpc"]
# total.AAV.samples <-
# total.AAV.samples[!grepl('4wks',total.AAV.samples$Group)]
transported.AAV.samples.30cpc <- total.AAV.samples[grepl("mRNA_30cpc_SN|mRNA_30cpc_Th|mRNA_30cpc_Ctx",
    total.AAV.samples$Group)]
transported.AAV.samples.3cpc <- total.AAV.samples[grepl("mRNA_3cpc_SN|30cpc_Th|mRNA_3cpc_Ctx",
    total.AAV.samples$Group)]
total.AAV.samples[, `:=`(Group, "mRNA_All")]
transported.AAV.samples.30cpc[, `:=`(Group, "mRNA_30cpc_Trsp")]
transported.AAV.samples.3cpc[, `:=`(Group, "mRNA_3cpc_Trsp")]

all.samples <- rbind(all.samples, total.AAV.samples, transported.AAV.samples.30cpc,
    transported.AAV.samples.3cpc)

rm(total.AAV.samples, transported.AAV.samples.30cpc, transported.AAV.samples.3cpc)

setkeyv(all.samples, c("Group", "Category", "GeneName", "structure", "start",
    "width", "Sequence", "seqlength"))

all.samples <- all.samples[, j = list(bitScore = sum(bitScore * tCount)/sum(tCount),
    mismatches = median(mismatches), mCount = sum(mCount), tCount = sum(tCount),
    BC = paste(unique(BC), collapse = ","), Animals = paste(unique(Sample),
        collapse = ","), LUTnrs = paste(unique(LUTnr), collapse = ","), RNAcount = sum(RNAcount),
    NormCount = log2(sum(RNAcount) + 1) * .N), by = c("Group", "Category", "GeneName",
    "structure", "start", "width", "Sequence", "seqlength")]

all.samples[, `:=`(start, floor((start + 2)/3))]
all.samples[, `:=`(width, ceiling((width)/3))]
all.samples[, `:=`(seqlength, ceiling(seqlength/3))]
all.samples[, `:=`(AA, floor(start + (width/2)))]
all.samples[, `:=`(AAproc, AA/seqlength * 100)]
```

Remove overhangs on the sequence based on the Structure annotation

```
all.samples[structure == "14aa", `:=`("Sequence", substr(Sequence, 3, 44))]
all.samples[structure == "22aa", `:=`("Sequence", substr(Sequence, 3, 68))]
all.samples[structure == "14aaG4S", `:=`("Sequence", substr(Sequence, 15, 56))]
all.samples[structure == "14aaA5", `:=`("Sequence", substr(Sequence, 15, 56))]

# Change the default behavior to induce start codons and Methionine
GENETIC_CODE_ALT <- GENETIC_CODE
attr(GENETIC_CODE_ALT, "alt_init_codons") <- c("TAA", "TAG")

all.samples[, `:=`(Peptide, mclapply(Sequence, function(x) as.character(Biostrings::translate(DNAString(x),
    genetic.code = GENETIC_CODE_ALT, if.fuzzy.codon = "solve")), mc.cores = detectCores()))]
```

```r
all.samples[, `:=`(Peptide, as.character(Peptide)), ]
saveRDS(all.samples, file = "data/allSamplesDataTable.RDS")

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

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

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

```
Time difference of 1.72911 hours
```

```r
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)
 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)
 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
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
multicore            * 0.2      2014-05-17 url
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
```