SCE assembly for the 2016-17 cohort

EJC

Feb 2021

Setup

The Babraham compute cluster does not contain a global tex installation, so a local tex is added to $PATH to allow knitting to pdf.

```
Sys.setenv(PATH=paste(Sys.getenv("PATH"),
                      "/bi/home/carre/texlive/2017/bin/x86_64-linux/",sep=":"))
```

Gwt counts

```
library(SingleCellExperiment)
```

```
## Loading required package: SummarizedExperiment
## Loading required package: GenomicRanges
## Loading required package: stats4
## Loading required package: BiocGenerics
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
## The following objects are masked from 'package:base':
##
##     Filter, Find, Map, Position, Reduce, anyDuplicated, append,
##     as.data.frame, basename, cbind, colnames, dirname, do.call,
##     duplicated, eval, evalq, get, grep, grepl, intersect, is.unsorted,
##     lapply, mapply, match, mget, order, paste, pmax, pmax.int, pmin,
##     pmin.int, rank, rbind, rownames, sapply, setdiff, sort, table,
##     tapply, union, unique, unsplit, which, which.max, which.min
```

```
## Loading required package: S4Vectors
##
## Attaching package: 'S4Vectors'
## The following object is masked from 'package:base':
##
##     expand.grid
## Loading required package: IRanges
## Loading required package: GenomeInfoDb
## Loading required package: Biobase
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
## Loading required package: DelayedArray
## Loading required package: matrixStats
##
## Attaching package: 'matrixStats'
## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians
## Loading required package: BiocParallel
##
## Attaching package: 'DelayedArray'
## The following objects are masked from 'package:matrixStats':
##
##     colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges
## The following objects are masked from 'package:base':
##
##     aperm, apply, rowsum
load("../data/SCE_incl_NTC.RData")  # this contains all raw counts, incl 8 NTCs.

sce <- sce[, !is.na(sce$PID)]
```

Make annotations the same as the GEO .xlsx submission file

```
GEO_compatible_names <- paste(sce$PID, sce$day, sce$age, sce$lib_plate,
    sce$lib_well, sep = "_")

# There is one library plate called 'lib' rather than 'cDNA':
GEO_compatible_names <- gsub(GEO_compatible_names, pattern = "lib",
    replacement = "cDNA")

############
```

```
count.matrix <- counts(sce)
colnames(count.matrix) <- GEO_compatible_names
```

Load in the GEO submission xlsx to check column names

```
library(openxlsx)

GEO <- read.xlsx("GEO submission.xlsx", sheet = 1, startRow = 22)

GEO <- GEO[1:952, ]  # the 10 x 96 well plates are covered in the 'samples' table here. (lower down is

#####

# Check both sets of names contain each other:
all(GEO$title %in% colnames(count.matrix))
```

```
## [1] TRUE
```

```
all(colnames(count.matrix) %in% GEO$title)
```

```
## [1] TRUE
```

```
###### Make the order the same:
ordered.count.matrix <- count.matrix[, GEO$title]

# check column names are identical:
all(colnames(ordered.count.matrix) == GEO$title)
```

```
## [1] TRUE
```

```
all(GEO$title == colnames(ordered.count.matrix))
```

```
## [1] TRUE
```

Write out supporting file

```
write.csv(ordered.count.matrix, file = "GEO_supporting_processed_data_file_raw_count_matrix.csv")

# ~ 110Mb file.

# Gzip:
system("gzip GEO_supporting_processed_data_file_raw_count_matrix.csv")
# ~ 7Mb file.
```

SessionInfo

```
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-pc-linux-gnu (64-bit)
```

```
## Running under: CentOS Linux 7 (Core)
##
## Matrix products: default
## BLAS:   /bi/apps/R/3.6.1/lib64/R/lib/libRblas.so
## LAPACK: /bi/apps/R/3.6.1/lib64/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C         LC_TIME=C
##  [4] LC_COLLATE=C         LC_MONETARY=C        LC_MESSAGES=C
##  [7] LC_PAPER=C           LC_NAME=C            LC_ADDRESS=C
## [10] LC_TELEPHONE=C       LC_MEASUREMENT=C     LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] openxlsx_4.1.4              SingleCellExperiment_1.8.0
##  [3] SummarizedExperiment_1.16.0 DelayedArray_0.12.0
##  [5] BiocParallel_1.20.0         matrixStats_0.55.0
##  [7] Biobase_2.46.0              GenomicRanges_1.38.0
##  [9] GenomeInfoDb_1.22.0         IRanges_2.20.1
## [11] S4Vectors_0.24.1            BiocGenerics_0.32.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3             knitr_1.26             XVector_0.26.0
##  [4] magrittr_1.5           zlibbioc_1.32.0        lattice_0.20-38
##  [7] rlang_0.4.10           stringr_1.4.0          tools_3.6.1
## [10] grid_3.6.1             xfun_0.11              htmltools_0.4.0
## [13] yaml_2.2.0             digest_0.6.23          zip_2.0.4
## [16] Matrix_1.2-17          GenomeInfoDbData_1.2.2 formatR_1.7
## [19] bitops_1.0-6           RCurl_1.95-4.12        evaluate_0.14
## [22] rmarkdown_2.0          stringi_1.4.3          compiler_3.6.1
```