mre.glmgampoi.bug.R

nathan

2022-03-30

library(SummarizedExperiment)
library(DelayedMatrixStats)
library(glmGamPoi)
library(HDF5Array)

################################################################
## read "small" test file as singlecellexperiment

sce = readRDS('sce.mre.rds')
test\_n = 10

smat = sce@assays@data$counts
metadata = sce@colData

################################################################
## create model matrix

metadata$c = metadata$seurat\_clusters
complex = with(metadata, interaction(c, Treatment) )
mm = model.matrix( ~ 0 + complex, data = metadata)
colnames(mm) = make.names(gsub('complex', 'cl', colnames(mm)))

################################################################
## check for empty interaction terms

mm.sub = mm[, apply(mm, 2, sum) > 10]
dim(mm.sub)

## [1] 489490 38

all.equal(colnames(smat), rownames(mm.sub))

## [1] TRUE

################################################################
## H5 settings
################################################################
## seems IMPORTANT for performance on large data sets
## unsure of how these settings interact with one another...
options(DelayedArray.block.size = 1e9)
setAutoBlockSize(1e9)

## automatic block size set to 1e+09 bytes (was 1e+09)

setAutoBlockShape('last-dim-grows-first')

## automatic block shape set to "last-dim-grows-first" (was "last-dim-grows-first")

## set h5 path and try to delete old objects
path = "./tnbc.h5"
thename = paste0('tnbc', 10 ) ; thename

## [1] "tnbc10"

thedims1 = paste0(".", thename, "\_dimnames/1"); thedims1

## [1] ".tnbc10\_dimnames/1"

thedims2 = paste0(".", thename, "\_dimnames/2"); thedims2

## [1] ".tnbc10\_dimnames/2"

try(h5delete(path, thename))
try(h5delete(path, thedims1) )
try(h5delete(path, thedims2) )
dim(smat)

## [1] 10 489490

################################################################
## write input data matrix as HDF5

smat.h5 = HDF5Array::writeHDF5Array(smat,
 path, name = thename,
 chunkdim = c(3, ncol(smat) ),
 verbose = TRUE, with.dimnames = TRUE, level = 6 )

## / Reading and realizing sparse block 1/1 ...

## OK

## \ Writing it ... OK

################################################################
## glmgampoi writes outputs to hdf5 too
## this is intended for the output
## managing h5 files is a pain, doesn't seem to be a global 'overwrite'
## I want to be able to find them later, so default R tmp spaces not great

tempdir = 'hdf5temp'
if( ! file.exists(tempdir) ) {
 unlink(tempdir, recursive = TRUE)
 dir.create(tempdir)
}
setHDF5DumpDir(dir = tempdir)
setHDF5DumpChunkLength(length=1000000000L)
setHDF5DumpChunkShape("last-dim-grows-first")
lsHDF5DumpFile()

## [1] group name otype dclass dim
## <0 rows> (or 0-length row.names)

################################################################
## build glmGamPoi model

lmf <- glm\_gp(
 smat.h5,
 mm.sub,
 on\_disk = TRUE,
 verbose = TRUE,
 offset = 0
)

## Calculate Size Factors (normed\_sum)
## You created a large dataset with compression and chunking.
## The chunk size is equal to the dataset dimensions.
## If you want to read subsets of the dataset, you should testsmaller chunk sizes to improve read times.
## Make initial dispersion estimate
## Make initial beta estimate
## Estimate beta
## You created a large dataset with compression and chunking.
## The chunk size is equal to the dataset dimensions.
## If you want to read subsets of the dataset, you should testsmaller chunk sizes to improve read times.
## You created a large dataset with compression and chunking.
## The chunk size is equal to the dataset dimensions.
## If you want to read subsets of the dataset, you should testsmaller chunk sizes to improve read times.
## Estimate dispersion
## Fit dispersion trend
## Shrink dispersion estimates
## Estimate beta again
## You created a large dataset with compression and chunking.
## The chunk size is equal to the dataset dimensions.
## If you want to read subsets of the dataset, you should testsmaller chunk sizes to improve read times.
## You created a large dataset with compression and chunking.
## The chunk size is equal to the dataset dimensions.
## If you want to read subsets of the dataset, you should testsmaller chunk sizes to improve read times.

str(lmf)

## List of 10
## $ Beta : num [1:10, 1:38] -6.679 -4.074 -7.065 -3.203 -0.988 ...
## ..- attr(\*, "dimnames")=List of 2
## .. ..$ : chr [1:10] "AL627309.1" "AL669831.5" "FAM87B" "LINC00115" ...
## .. ..$ : chr [1:38] "cl0.Anti.PD.L1.Chemo" "cl1.Anti.PD.L1.Chemo" "cl2.Anti.PD.L1.Chemo" "cl3.Anti.PD.L1.Chemo" ...
## $ overdispersions : Named num [1:10] 93.853 13.606 178.115 4.961 0.262 ...
## ..- attr(\*, "names")= chr [1:10] "AL627309.1" "AL669831.5" "FAM87B" "LINC00115" ...
## $ overdispersion\_shrinkage\_list:List of 5
## ..$ dispersion\_trend: num [1:10] 6.66 6.66 6.66 6.66 6.66 ...
## ..$ ql\_disp\_estimate: num [1:10] 1.045 1.035 1.048 0.981 0.58 ...
## ..$ ql\_disp\_trend : num [1:10] 0.663 0.663 0.663 0.663 0.663 ...
## ..$ ql\_disp\_shrunken: num [1:10] 1.045 1.035 1.048 0.981 0.58 ...
## ..$ ql\_df0 : num 6
## $ deviances : Named num [1:10] 2742 16081 1582 28508 67565 ...
## ..- attr(\*, "names")= chr [1:10] "AL627309.1" "AL669831.5" "FAM87B" "LINC00115" ...
## $ Mu :Formal class 'DelayedMatrix' [package "DelayedArray"] with 1 slot
## .. ..@ seed:Formal class 'DelayedSetDimnames' [package "DelayedArray"] with 2 slots
## .. .. .. ..@ dimnames:List of 2
## .. .. .. .. ..$ : chr [1:10] "AL627309.1" "AL669831.5" "FAM87B" "LINC00115" ...
## .. .. .. .. ..$ : chr [1:489490] "AAACCTGAGGTTACCT.Pre\_P007\_b" "AAACCTGCAAAGGAAG.Pre\_P007\_b" "AAACCTGCAAGTCTAC.Pre\_P007\_b" "AAACCTGCAATAAGCA.Pre\_P007\_b" ...
## .. .. .. ..@ seed :Formal class 'HDF5ArraySeed' [package "HDF5Array"] with 7 slots
## .. .. .. .. .. ..@ filepath : chr "/Users/nathan/Documents/R/ZhangTNBC/hdf5temp/auto00020.h5"
## .. .. .. .. .. ..@ name : chr "/HDF5ArrayAUTO00020"
## .. .. .. .. .. ..@ as\_sparse: logi FALSE
## .. .. .. .. .. ..@ type : chr NA
## .. .. .. .. .. ..@ dim : int [1:2] 10 489490
## .. .. .. .. .. ..@ chunkdim : int [1:2] 10 489490
## .. .. .. .. .. ..@ first\_val: num 1.26e-06
## $ size\_factors : Named num [1:489490] 0.001 0.001 0.001 0.001 1.197 ...
## ..- attr(\*, "names")= chr [1:489490] "AAACCTGAGGTTACCT.Pre\_P007\_b" "AAACCTGCAAAGGAAG.Pre\_P007\_b" "AAACCTGCAAGTCTAC.Pre\_P007\_b" "AAACCTGCAATAAGCA.Pre\_P007\_b" ...
## $ Offset :Formal class 'DelayedMatrix' [package "DelayedArray"] with 1 slot
## .. ..@ seed:Formal class 'DelayedSetDimnames' [package "DelayedArray"] with 2 slots
## .. .. .. ..@ dimnames:List of 2
## .. .. .. .. ..$ : chr [1:10] "AL627309.1" "AL669831.5" "FAM87B" "LINC00115" ...
## .. .. .. .. ..$ : chr [1:489490] "AAACCTGAGGTTACCT.Pre\_P007\_b" "AAACCTGCAAAGGAAG.Pre\_P007\_b" "AAACCTGCAAGTCTAC.Pre\_P007\_b" "AAACCTGCAATAAGCA.Pre\_P007\_b" ...
## .. .. .. ..@ seed :Formal class 'HDF5ArraySeed' [package "HDF5Array"] with 7 slots
## .. .. .. .. .. ..@ filepath : chr "/Users/nathan/Documents/R/ZhangTNBC/hdf5temp/auto00016.h5"
## .. .. .. .. .. ..@ name : chr "/HDF5ArrayAUTO00016"
## .. .. .. .. .. ..@ as\_sparse: logi FALSE
## .. .. .. .. .. ..@ type : chr NA
## .. .. .. .. .. ..@ dim : int [1:2] 10 489490
## .. .. .. .. .. ..@ chunkdim : int [1:2] 10 489490
## .. .. .. .. .. ..@ first\_val: num -6.91
## $ ridge\_penalty : Named num [1:38] 2.04e-16 2.04e-16 2.04e-16 2.04e-16 2.04e-16 ...
## ..- attr(\*, "target")= num [1:38] 0 0 0 0 0 0 0 0 0 0 ...
## ..- attr(\*, "names")= chr [1:38] "cl0.Anti.PD.L1.Chemo" "cl1.Anti.PD.L1.Chemo" "cl2.Anti.PD.L1.Chemo" "cl3.Anti.PD.L1.Chemo" ...
## $ data :Formal class 'SummarizedExperiment' [package "SummarizedExperiment"] with 5 slots
## .. ..@ colData :Formal class 'DFrame' [package "S4Vectors"] with 6 slots
## .. .. .. ..@ rownames : chr [1:489490] "AAACCTGAGGTTACCT.Pre\_P007\_b" "AAACCTGCAAAGGAAG.Pre\_P007\_b" "AAACCTGCAAGTCTAC.Pre\_P007\_b" "AAACCTGCAATAAGCA.Pre\_P007\_b" ...
## .. .. .. ..@ nrows : int 489490
## .. .. .. ..@ listData : Named list()
## .. .. .. ..@ elementType : chr "ANY"
## .. .. .. ..@ elementMetadata: NULL
## .. .. .. ..@ metadata : list()
## .. ..@ assays :Formal class 'SimpleAssays' [package "SummarizedExperiment"] with 1 slot
## .. .. .. ..@ data:Formal class 'SimpleList' [package "S4Vectors"] with 4 slots
## .. .. .. .. .. ..@ listData :List of 1
## .. .. .. .. .. .. ..$ counts:Formal class 'HDF5Matrix' [package "HDF5Array"] with 1 slot
## .. .. .. .. .. .. .. .. ..@ seed:Formal class 'HDF5ArraySeed' [package "HDF5Array"] with 7 slots
## .. .. .. .. .. .. .. .. .. .. ..@ filepath : chr "/Users/nathan/Documents/R/ZhangTNBC/tnbc.h5"
## .. .. .. .. .. .. .. .. .. .. ..@ name : chr "/tnbc10"
## .. .. .. .. .. .. .. .. .. .. ..@ as\_sparse: logi TRUE
## .. .. .. .. .. .. .. .. .. .. ..@ type : chr NA
## .. .. .. .. .. .. .. .. .. .. ..@ dim : int [1:2] 10 489490
## .. .. .. .. .. .. .. .. .. .. ..@ chunkdim : int [1:2] 3 489490
## .. .. .. .. .. .. .. .. .. .. ..@ first\_val: num 0
## .. .. .. .. .. ..@ elementType : chr "ANY"
## .. .. .. .. .. ..@ elementMetadata: NULL
## .. .. .. .. .. ..@ metadata : list()
## .. ..@ NAMES : chr [1:10] "AL627309.1" "AL669831.5" "FAM87B" "LINC00115" ...
## .. ..@ elementMetadata:Formal class 'DFrame' [package "S4Vectors"] with 6 slots
## .. .. .. ..@ rownames : NULL
## .. .. .. ..@ nrows : int 10
## .. .. .. ..@ listData : Named list()
## .. .. .. ..@ elementType : chr "ANY"
## .. .. .. ..@ elementMetadata: NULL
## .. .. .. ..@ metadata : list()
## .. ..@ metadata : list()
## $ model\_matrix : num [1:489490, 1:38] 1 0 0 0 1 0 0 1 1 0 ...
## ..- attr(\*, "dimnames")=List of 2
## .. ..$ : chr [1:489490] "AAACCTGAGGTTACCT.Pre\_P007\_b" "AAACCTGCAAAGGAAG.Pre\_P007\_b" "AAACCTGCAAGTCTAC.Pre\_P007\_b" "AAACCTGCAATAAGCA.Pre\_P007\_b" ...
## .. ..$ : chr [1:38] "cl0.Anti.PD.L1.Chemo" "cl1.Anti.PD.L1.Chemo" "cl2.Anti.PD.L1.Chemo" "cl3.Anti.PD.L1.Chemo" ...
## ..- attr(\*, "intercept\_position")= num 0
## - attr(\*, "class")= chr "glmGamPoi"

################################################################
## make a contrast vector

con.treat = scale( scale = FALSE, as.integer(grepl( "L1", colnames(lmf$Beta) ) ) ) \* 2
table(con.treat)

## con.treat
## -1 1
## 19 19

################################################################
## perform a test

compare.treat = try(test\_de( lmf, con.treat, verbose = TRUE ))

## Fit reduced model

## Error in combine\_size\_factors\_and\_offset(offset, size\_factors, Y, verbose = verbose) :
## length(offset) == 1 || length(offset) == n\_samples is not TRUE

################################################################
## take the h5 matrix and realize it
## build model and run again with on\_disk = FALSE

smat.realized = as.matrix(smat.h5)

lmf2 <- glm\_gp(
 smat.realized,
 mm.sub,
 on\_disk = FALSE,
 verbose = TRUE,
 offset = 0
)

## Calculate Size Factors (normed\_sum)
## Make initial dispersion estimate
## Make initial beta estimate
## Estimate beta
## Estimate dispersion
## Fit dispersion trend
## Shrink dispersion estimates
## Estimate beta again

################################################################
## make contrasts and run test\_de on realized matrix and model

con.treat = scale( scale = FALSE, as.integer(grepl( "L1", colnames(lmf2$Beta) ) ) ) \* 2
compare.treat = test\_de( lmf2, con.treat, verbose = TRUE )

## Fit reduced model
## Calculate quasi likelihood ratio
## Prepare results

head(compare.treat)

## name pval adj\_pval f\_statistic df1 df2 lfc
## 1 AL627309.1 9.973956e-01 9.982485e-01 1.065449e-05 1 489458 4.328085e+08
## 2 AL669831.5 7.292061e-01 9.982485e-01 1.198419e-01 1 489458 1.203745e+00
## 3 FAM87B 9.982485e-01 9.982485e-01 4.818658e-06 1 489458 4.328085e+08
## 4 LINC00115 1.136327e-01 2.840818e-01 2.502967e+00 1 489458 3.920252e+00
## 5 NOC2L 3.231114e-17 3.231114e-16 7.120328e+01 1 489458 7.272998e+00
## 6 KLHL17 7.932391e-01 9.982485e-01 6.870004e-02 1 489458 -1.442695e+08

## session information
sessionInfo()

## R version 4.1.2 (2021-11-01)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Monterey 12.3
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8
##
## attached base packages:
## [1] stats4 stats graphics grDevices utils datasets methods
## [8] base
##
## other attached packages:
## [1] Seurat\_4.1.0 SeuratObject\_4.0.4
## [3] HDF5Array\_1.22.1 rhdf5\_2.38.1
## [5] DelayedMatrixStats\_1.16.0 DelayedArray\_0.20.0
## [7] Matrix\_1.4-1 SummarizedExperiment\_1.24.0
## [9] Biobase\_2.54.0 GenomicRanges\_1.46.1
## [11] GenomeInfoDb\_1.30.1 IRanges\_2.28.0
## [13] S4Vectors\_0.32.4 BiocGenerics\_0.40.0
## [15] MatrixGenerics\_1.6.0 matrixStats\_0.61.0
## [17] glmGamPoi\_1.6.0
##
## loaded via a namespace (and not attached):
## [1] Rtsne\_0.15 colorspace\_2.0-3
## [3] deldir\_1.0-6 ellipsis\_0.3.2
## [5] ggridges\_0.5.3 XVector\_0.34.0
## [7] spatstat.data\_2.1-4 leiden\_0.3.9
## [9] listenv\_0.8.0 ggrepel\_0.9.1
## [11] fansi\_1.0.3 codetools\_0.2-18
## [13] splines\_4.1.2 sparseMatrixStats\_1.6.0
## [15] knitr\_1.38 polyclip\_1.10-0
## [17] jsonlite\_1.8.0 ica\_1.0-2
## [19] cluster\_2.1.3 png\_0.1-7
## [21] uwot\_0.1.11 shiny\_1.7.1
## [23] sctransform\_0.3.3 spatstat.sparse\_2.1-0
## [25] compiler\_4.1.2 httr\_1.4.2
## [27] assertthat\_0.2.1 fastmap\_1.1.0
## [29] lazyeval\_0.2.2 cli\_3.2.0
## [31] later\_1.3.0 htmltools\_0.5.2
## [33] tools\_4.1.2 igraph\_1.2.11
## [35] gtable\_0.3.0 glue\_1.6.2
## [37] GenomeInfoDbData\_1.2.7 RANN\_2.6.1
## [39] reshape2\_1.4.4 dplyr\_1.0.8
## [41] Rcpp\_1.0.8.3 scattermore\_0.8
## [43] vctrs\_0.3.8 rhdf5filters\_1.6.0
## [45] nlme\_3.1-157 lmtest\_0.9-40
## [47] spatstat.random\_2.1-0 xfun\_0.30
## [49] stringr\_1.4.0 globals\_0.14.0
## [51] beachmat\_2.10.0 mime\_0.12
## [53] miniUI\_0.1.1.1 lifecycle\_1.0.1
## [55] irlba\_2.3.5 goftest\_1.2-3
## [57] future\_1.24.0 zlibbioc\_1.40.0
## [59] MASS\_7.3-56 zoo\_1.8-9
## [61] scales\_1.1.1 spatstat.core\_2.4-0
## [63] promises\_1.2.0.1 spatstat.utils\_2.3-0
## [65] parallel\_4.1.2 RColorBrewer\_1.1-2
## [67] yaml\_2.3.5 SingleCellExperiment\_1.16.0
## [69] reticulate\_1.24 pbapply\_1.5-0
## [71] gridExtra\_2.3 ggplot2\_3.3.5
## [73] rpart\_4.1.16 stringi\_1.7.6
## [75] highr\_0.9 rlang\_1.0.2
## [77] pkgconfig\_2.0.3 bitops\_1.0-7
## [79] evaluate\_0.15 lattice\_0.20-45
## [81] tensor\_1.5 ROCR\_1.0-11
## [83] purrr\_0.3.4 Rhdf5lib\_1.16.0
## [85] patchwork\_1.1.1 htmlwidgets\_1.5.4
## [87] cowplot\_1.1.1 tidyselect\_1.1.2
## [89] parallelly\_1.30.0 RcppAnnoy\_0.0.19
## [91] plyr\_1.8.7 magrittr\_2.0.2
## [93] R6\_2.5.1 generics\_0.1.2
## [95] DBI\_1.1.2 mgcv\_1.8-40
## [97] pillar\_1.7.0 fitdistrplus\_1.1-8
## [99] abind\_1.4-5 survival\_3.3-1
## [101] RCurl\_1.98-1.6 tibble\_3.1.6
## [103] future.apply\_1.8.1 crayon\_1.5.1
## [105] KernSmooth\_2.23-20 utf8\_1.2.2
## [107] spatstat.geom\_2.4-0 plotly\_4.10.0
## [109] rmarkdown\_2.13 grid\_4.1.2
## [111] data.table\_1.14.2 digest\_0.6.29
## [113] xtable\_1.8-4 tidyr\_1.2.0
## [115] httpuv\_1.6.5 munsell\_0.5.0
## [117] viridisLite\_0.4.0