preprocessing

Matthew Angel

7/28/2021

MSD Preprocessing

```
suppressMessages(library(stringr))
suppressMessages(library(tidyverse))

workDir <- '/Users/angelmg/Documents/nci_vb_git/bergamaschi_pfizer_cancer'
setwd(workDir)

raw_data <- read.delim('data/cancer_deidentified.csv', header = TRUE, sep=',',check.names = FALSE)

#Elimitate C30
raw_data <- raw_data %>% filter(!grepl("^C30_", Vial_Label))

row.names(raw_data) <- raw_data$Vial_Label
raw_data$Vial_Label <- NULL

raw_data <- t(as.matrix(raw_data))

class(raw_data) <- "numeric"

df <- log2(raw_data + 1)

# Setup metadata
samples <- colnames(df)

patient_id <- apply(array(samples), 1, function(z) unlist(str_split(z,"_"))[1])
timepoint <- apply(array(samples), 1, function(z) unlist(str_split(z,"_"))[2])
timepoint <- gsub("D","d",timepoint)

annot <- data.frame(sample_id = samples, patient_id = patient_id, timepoint = timepoint)

# Do diff counts
contrasts <- c("d2-d1","d23-d22","d23-d22-d2-d1")

# Setup annotation metadata
contrast_samples <- apply(expand.grid(unique(patient_id), contrasts), 1, paste, collapse="_")
annot_diff <- data.frame(sample_id = contrast_samples)
annot_diff$patient_id <- rep(unique(patient_id),length(contrasts))
annot_diff$timepoint <- apply(array(annot_diff$sample_id), 1, function(z) unlist(str_split(z,"_"))[2])

for(i in seq_along(contrasts)){
  contrast <- contrasts[i]
```

```
  #contrast <- contrasts[3]

  if(length(unlist(str_split(contrast,"-"))) == 2){
    test <- unlist(str_split(contrast,"-"))[1]
    ref  <- unlist(str_split(contrast,"-"))[2]

    annot.c <- annot %>% filter(timepoint %in% c(test,ref)) %>% arrange(patient_id)
    complex <- FALSE
  }else{
    test <- paste(unlist(str_split(contrast,"-"))[1:2],collapse="-")
    ref  <- paste(unlist(str_split(contrast,"-"))[3:4],collapse="-")

    annot.c <- annot_diff %>% filter(timepoint %in% c(test,ref)) %>% arrange(patient_id)
    complex <- TRUE
  }

  test_samples <- annot.c$sample_id[annot.c$timepoint == test]
  ref_samples <- annot.c$sample_id[annot.c$timepoint == ref]

  test_animals <- gsub(paste0("_",test),"",test_samples, ignore.case = TRUE)
  ref_animals <- gsub(paste0("_",ref),"",ref_samples, ignore.case = TRUE)

  if(!all(test_animals == ref_animals)){
    stop("Something wrong with animal order")
  }

  samples_in_contrast <- annot.c$sample_id

  if(!complex){
    df.test <- df[,test_samples]
    df.ref  <- df[,ref_samples]
  }else{
    df.test <- df_diff[,test_samples]
    df.ref  <- df_diff[,ref_samples]
  }

  df.ret <- df.test - df.ref
  colnames(df.ret) <- paste(test_animals,contrast,sep="_")

  if(i == 1){
    df_diff <- df.ret
  }else{
    df_diff <- cbind(df_diff,df.ret)
  }

}

colnames(annot_diff)[which(colnames(annot_diff)=="timepoint")] <- "contrast" #need to add contrast here

write.table(df, file = file.path(workDir,"output","processed_data.csv"), row.names = TRUE, quote = FALSE
write.table(annot, file = file.path(workDir,"output","sample_annot.csv"), row.names = TRUE, quote = FALS
write.table(df_diff, file = file.path(workDir,"output","diff_counts.csv"), row.names = TRUE, quote = FA
write.table(annot_diff, file = file.path(workDir,"output","annot_diff.csv"), row.names = FALSE, quote =
```

Session Info

```
sessionInfo()
```

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.7
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/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] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] forcats_0.5.1   dplyr_1.0.7     purrr_0.3.4     readr_2.1.0
## [5] tidyr_1.1.4     tibble_3.1.6    ggplot2_3.3.5   tidyverse_1.3.1
## [9] stringr_1.4.0
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.1    xfun_0.28           haven_2.4.3
##  [4] colorspace_2.0-2    vctrs_0.3.8         generics_0.1.1
##  [7] htmltools_0.5.2     yaml_2.2.1          utf8_1.2.2
## [10] rlang_0.4.12        pillar_1.6.4        glue_1.5.0
## [13] withr_2.4.2         DBI_1.1.1           dbplyr_2.1.1
## [16] modelr_0.1.8        readxl_1.3.1        lifecycle_1.0.1
## [19] munsell_0.5.0       gtable_0.3.0        cellranger_1.1.0
## [22] rvest_1.0.2         evaluate_0.14       knitr_1.36
## [25] tzdb_0.2.0          fastmap_1.1.0       fansi_0.5.0
## [28] broom_0.7.10        Rcpp_1.0.7          scales_1.1.1
## [31] backports_1.3.0     BiocManager_1.30.16 jsonlite_1.7.2
## [34] fs_1.5.0            hms_1.1.1           digest_0.6.28
## [37] stringi_1.7.5       grid_4.0.5          cli_3.1.0
## [40] tools_4.0.5         magrittr_2.0.1      crayon_1.4.2
## [43] pkgconfig_2.0.3     ellipsis_0.3.2      xml2_1.3.2
## [46] reprex_2.0.1        lubridate_1.8.0     rstudioapi_0.13
## [49] assertthat_0.2.1    rmarkdown_2.11      httr_1.4.2
## [52] R6_2.5.1            compiler_4.0.5
```