AllData_integration_preQC

December 25, 2025

1 function for QC threshold calculation

[ ]:

```
getThreshold <- function(x, iqr.multiplier = 3, only.high = TRUE){
                    x.med <- median(x)
                    outs <- boxplot.stats(x, coef = iqr.multiplier)$out
                    if(only.high){
                        Threshold <- subset(outs, outs > x.med)
                    } else {
                        Threshold <- outs
                    }
                    return(Threshold)
                }
```

2 cell QC

measurement: nFeature_RNA, nCount_RNA, percent.mt

[ ]:

```
objList <- list.files('/project/sex_cancer/data/step2_standardization', pattern␣
 ↪= 'obj', full.names = TRUE)
objList
length(objList)
```

[ ]:

```
qcList <- lapply(objList, function(x){
                    obj <- readRDS(x) %>% PercentageFeatureSet(pattern =␣
 ↪"^MT-", col.name = "percent.mt")
                    return(obj@meta.data)
              })
metadata_cellQC <- qcList %>% do.call(rbind, .)
metadata_cellQC %>% head(n = 2)
```

[ ]:

```
## threshold calculation
nFeature_thres <- getThreshold(metadata_cellQC$nFeature_RNA, iqr.multiplier =␣
 ↪3) %>% min()
nFeature_thres
```

```
nCount_thres <- getThreshold(metadata_cellQC$nCount_RNA, iqr.multiplier = 3)␣
 ↪%>% min()
nCount_thres
```

[ ]:

```
## assign cell class (outlier or not)
metadata_cellQC <- metadata_cellQC %>%
                   mutate(Class = case_when((nFeature_RNA >= nFeature_thres |␣
 ↪nCount_RNA >= nCount_thres | percent.mt > 40) ~ "Outlier",
                                            TRUE ~ "Keep"))
```

2.1 cell QC statistics

[ ]:

```
table(metadata_cellQC$Cohort, metadata_cellQC$Class) %>% as.data.frame.matrix()␣
 ↪%>% mutate(ratio_outlier = Keep/(Keep+Outlier)*100)
# arrange(ratio_outlier)
```

2.2 filter cell

[ ]:

```
metadata_cellQC2 <- metadata_cellQC %>%
                    subset(Class == "Keep")
metadata_cellQC2 %$% table(.$Cohort) %>% as.data.frame() %>% subset(Freq>0) %>%␣
 ↪arrange(desc(Freq))
```

3 sample QC

measurement: cell number

[ ]:

```
## sample-level statistics
metadata_sampleQC <- metadata_cellQC2 %>%
                     group_by(Cohort, SampleID, SampleType, Sex) %>%
                     summarize(Ncell = n(), .groups = 'drop')
metadata_sampleQC
```

[ ]:

```
metadata_sampleQC2 <- metadata_sampleQC %>%
                      subset(Ncell >= 100)
metadata_sampleQC %>%
subset(Ncell < 100) %>%
group_by(Cohort, SampleType, Sex) %>%
summarize(Nsample = n(), .groups = 'drop')
```

4 perform QC

[ ]:

```
metadata_keep <- metadata_cellQC2 %>%
                 subset(SampleID %in% metadata_sampleQC2$SampleID) ## discard␣
 ↪samples with <= 100 cells
metadata_keep %>% head(n = 2)
```

[ ]:

```
objList <- list.files('/project/sex_cancer/data/step2_standardization', pattern␣
 ↪= 'obj', full.names = TRUE)
objList
length(objList)
```

[ ]:

```
## filter and save
lapply(objList, function(x){
    print(x)
    obj <- readRDS(x)
    cell_keep <- intersect(rownames(metadata_keep), colnames(obj))
    obj_new <- obj %>% subset(cells = cell_keep)
    saveRDS(obj_new, gsub("step2_standardization", "step3_integration", x))
})
```

5 SexTumorDB statistics

[ ]:

```
objList2 <- list.files("/project/sex_cancer/data/step3_integration", pattern =␣
 ↪'obj', full.names = TRUE)
objList2
length(objList2)
```

[ ]:

```
metaList <- lapply(objList2, function(x){
                     obj <- readRDS(x)
                     obj@meta.data
           })
length(metaList)
```

5.1 cell statistics

[ ]:

```
meta_cell <- metaList %>% do.call(rbind, .)
meta_cell <- meta_cell %>% mutate_if(~!is.numeric(.), ext_list)

dim(meta_cell) ## 2,014,043 cells
meta_cell %>% head(n = 2)
```

[ ]:

```
saveRDS(meta_cell, "/project/sex_cancer/data/step3_integration/metadata_cell.
 ↪rds")
write.csv(meta_cell, "/project/sex_cancer/data/step3_integration/metadata_cell.
 ↪csv", row.names = FALSE, quote = FALSE)
```

5.2 sample statistics

```
[ ]: meta_sample <- meta_cell %>% dplyr::select(c("Cohort", "SampleID",␣
      ↪"SampleType", "DonorID", "Sex", "Chemistry", "Tissue")) %>% .[!duplicated(.
      ↪$SampleID),]
     rownames(meta_sample) <- NULL

     dim(meta_sample) ## 532 samples
     meta_sample %>% head(n = 2)
```

```
[ ]: saveRDS(meta_sample, "/project/sex_cancer/data/step3_integration/
      ↪metadata_sample.rds")
     write.csv(meta_sample, "/project/sex_cancer/data/step3_integration/
      ↪metadata_sample.csv", row.names = FALSE, quote = FALSE)
```

5.3 Cohort statistics

```
[ ]: meta_cohort <- merge(meta_cell %>% group_by(Cohort, SampleType, Sex) %>%␣
      ↪summarize(Ncell = n(), .groups = 'drop'),
                          meta_sample %>% group_by(Cohort, SampleType, Sex) %>%␣
      ↪summarize(Nsample = n(), .groups = 'drop'),
                          by = c("Cohort", "SampleType", "Sex"), alll = TRUE) %>%
                    mutate(SampleType = factor(SampleType, levels = c("tumor",␣
      ↪"normal", "normal_adjacent"))) %>%
                    arrange(Cohort, SampleType, Sex)
     meta_cohort ## 40 combinations
```

```
[ ]: write.csv(meta_cohort, "/project/sex_cancer/data/step3_integration/
      ↪metadata_cohort.csv", row.names = FALSE, quote = FALSE)
     saveRDS(meta_cohort, "/project/sex_cancer/data/step3_integration/
      ↪metadata_cohort.rds")
```