THCA_Pu2021_process

December 25, 2025

1 load data

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE184362 - all without treatment

[ ]:
```
obj.THCA <- readRDS('/project/sex_cancer/data/THCA_Pu2021/AllComb/expr.RDS')

info <- read_xlsx('/project/sex_cancer/data/THCA_Pu2021/THCA_PatientInfo.xlsx')
colnames(info)[c(2,4,5)] <- c('sampleID','ID2','sample.ident')

meta <- obj.THCA@meta.data %>%
        transform(sample.ident = gsub('-','_',meta$sample.ident)) %>%
        separate('sample.ident',into = c('sampleID','other'),sep = '_',remove =␣
 ↪FALSE)  %>%
        merge(., info[,c(2:3)],by = 'sampleID',all.x = TRUE)
rownames(meta) <- meta$barcode
obj.THCA@meta.data <- meta
```

2 modify meta.data

[ ]:
```
meta_drop <- c(names(obj.THCA@meta.data) %>% .[grepl(".corr", .)], names(obj.
 ↪THCA@meta.data) %>% .[grepl("GS__", .)])
obj.THCA@meta.data <- obj.THCA@meta.data %>% dplyr::select(-meta_drop)

sample_paratumor <- unique(obj.THCA@meta.data$orig.ident) %>% .[grepl("_P", .)]
sample_tumor <- unique(obj.THCA@meta.data$orig.ident) %>% .[grepl("_T", .)]
```

[ ]:
```
obj.THCA@meta.data <- obj.THCA@meta.data %>%
                      dplyr::select(-c('diss.percent', 'RNA_snn_res.0.8',␣
 ↪'seurat_clusters', 'CellCycle.score', 'Malign.type', 'Malign.score',
                                       'Stemness.score', 'orig.ident',␣
 ↪'default', 'other', 'integrated_snn_res.0.8', 'comb.cluster', 'Cell.Type'))␣
 ↪%>%
                      dplyr::rename(c('SampleID' = 'sample.ident', 'DonorID' =␣
 ↪'sampleID')) %>%
                      transform(Cohort = 'THCA_Pu2021')  %>%
```

```
                    mutate(Sex = case_when(Sex == 'female' ~ 'F', Sex ==␣
↪'male' ~ 'M', TRUE ~ 'Others')) %>%
                    mutate(SampleType = case_when(SampleID %in%␣
↪sample_paratumor ~ 'normal_adjacent', SampleID %in% sample_tumor ~ 'tumor',␣
↪TRUE ~ 'Others'))
```

3 cell type annotation

[ ]:
```
marker_list <- c('PTPRC',
                 'LYZ', 'S100A8', 'S100A9', 'CD14', 'CD68', 'CD163', 'CD1C',␣
↪'LAMP3', 'CSF3R', 'CLEC10A', ## Myeloid 'TPSAB1',
                 'CD3D', 'CD3E', 'CD8A', 'CD4', 'TRAC', 'TRDC', 'IL7R',␣
↪'CTLA4', 'FOXP3', 'NKG7', 'NCAM1', 'KLRD1', ## T/NK
                 'CD79A', 'CD79B', 'MS4A1', 'IGHM', 'IGHD', 'IGKC', 'CD74', ##␣
↪B plasma
                 'TG', 'EPCAM', 'KRT18', 'KRT19', 'CLU', 'FN1', 'MGST1',␣
↪'S100A13', ## thyrocyte=thyroid epithelial cells
                 'COL1A1', 'COL1A2', 'ACTA2', ## Fibro
                 'PECAM1', 'CDH5', 'VWF' ## Endo
                )
```

[ ]:
```
sampleList = unique(ext_list(obj.THCA$SampleID))
obj.anchor <- lapply(sampleList, function(sampleID){
                    obj <- obj.THCA %>%
                           subset(SampleID == sampleID) %>%
                           NormalizeData(normalization.method =␣
↪"LogNormalize", scale.factor = 10000, verbose = F) %>%
                           FindVariableFeatures(selection.method = "vst",␣
↪nfeatures = 3000, verbose = F)
                    VariableFeatures(obj) <- union(marker_list,␣
↪VariableFeatures(obj))
                    obj <- obj %>% ScaleData(vars.to.regress =␣
↪c("nCount_RNA", "percent.mt"),verbose = F)
                    return(obj)
                  })
names(obj.anchor) <- sampleList
## FindIntegrationAnchors
obj.anchor <- FindIntegrationAnchors(obj.anchor, dims = 1:30)
obj.anchor <- IntegrateData(anchorset = obj.anchor, dims = 1:30, verbose = F)

## scale data+runPCA
obj.anchor <- obj.anchor %>%
              ScaleData(verbose = FALSE) %>%
              RunPCA(npcs = 50, verbose = F)
```

```
#### Clustering
set.seed(486)
select <- 1:(PC_selection(obj.anchor)$PCselect %>% min())
obj.anchor <- obj.anchor %>%
              RunUMAP(reduction = "pca", dims = select, umap.method = "uwot")␣
 ↪%>%
              RunTSNE(reduction = "pca", dims = select) %>%
              FindNeighbors(reduction = "pca", dims = select) %>%
              FindClusters(resolution = 0.1) %>% FindClusters(resolution = 0.
 ↪2) %>% FindClusters(resolution = 0.3) %>% FindClusters(resolution = 0.4)
colnames(obj.anchor@meta.data) <- colnames(obj.anchor@meta.data) %>%␣
 ↪gsub("integrated_snn_res.0.", "r0", .)
```

3.1 assign mCT

[ ]:
```
obj.THCA <- obj.anchor
obj.THCA@meta.data <- obj.THCA@meta.data %>% dplyr::
 ↪select(-c('seurat_clusters', 'r01', 'r03', 'r04'))
obj.THCA@meta.data <- obj.THCA@meta.data %>%
                      mutate(mCT = case_when(r02 %in% c('1') ~ 'Epi', ##␣
 ↪thyrocyte=thyroid epithelial cells
                                             r02 %in% c('6') ~ 'Myeloid',
                                             r02 %in% c('2', '5') ~ 'B/Plasma',
                                             r02 %in% c('0', '3', '4', '7') ~␣
 ↪'T/NK',
                                             r02 %in% c('8') ~ 'Endo',
                                             r02 %in% c('9') ~ 'Fibro',
                                             TRUE ~ 'Others'
                                             ))
```

3.2 assign NK/T/Epi

[ ]:
```
obj.THCA.nkte <- obj.THCA %>% subset(mCT %in% c('T/NK', 'Epi'))
DefaultAssay(obj.THCA.nkte) <- 'integrated'

obj.THCA.nkte <- obj.THCA.nkte %>%
                 FindVariableFeatures(selection.method = "vst", nfeatures =␣
 ↪1500, verbose = F) %>%
                 ScaleData(verbose = FALSE) %>%
                 RunPCA(npcs = 50, verbose = F)
select <- 1:(PC_selection(obj.THCA.nkte)$PCselect %>% min())
obj.THCA.nkte <- obj.THCA.nkte %>%
                 RunUMAP(reduction = "pca", dims = select, umap.method =␣
 ↪"uwot") %>%
                 RunTSNE(reduction = "pca", dims = select) %>%
                 FindNeighbors(reduction = "pca", dims = select) %>%
```

```
                  FindClusters(resolution = 0.1) %>% FindClusters(resolution = 0.
 ↪2) %>% FindClusters(resolution = 0.3) %>% FindClusters(resolution = 0.4)
colnames(obj.THCA.nkte@meta.data) <- colnames(obj.THCA.nkte@meta.data) %>%␣
 ↪gsub("integrated_snn_res.0.", "nkte0", .)
```

[ ]:
```
obj.THCA.epi <- obj.THCA.nkte %>% subset(nkte01 == 0)
obj.THCA.epi@meta.data <- obj.THCA.epi@meta.data %>%
                          transform(dCT = 'Epi') %>%
                          dplyr::select(-c('nkte01', 'nkte02', 'nkte03',␣
 ↪'nkte04', 'seurat_clusters', 'r02'))
obj.THCA.epi@meta.data %>% head(n = 2)
```

3.2.1 assign NK/T

[ ]:
```
obj.THCA.nkt <- obj.THCA.nkte %>% subset(nkte01 != 0)
DefaultAssay(obj.THCA.nkt) <- 'integrated'
## scale data+runPCA
obj.THCA.nkt <- obj.THCA.nkt %>%
                  FindVariableFeatures(selection.method = "vst", nfeatures =␣
 ↪1500, verbose = F) %>%
                  ScaleData(verbose = FALSE) %>%
                  RunPCA(npcs = 50, verbose = F)
select <- 1:(PC_selection(obj.THCA.nkt)$PCselect %>% min())
obj.THCA.nkt <- obj.THCA.nkt %>%
                  RunUMAP(reduction = "pca", dims = select, umap.method =␣
 ↪"uwot") %>%
                  RunTSNE(reduction = "pca", dims = select) %>%
                  FindNeighbors(reduction = "pca", dims = select) %>%
                  FindClusters(resolution = 0.1) %>% FindClusters(resolution = 0.
 ↪2) %>% FindClusters(resolution = 0.3) %>% FindClusters(resolution = 0.4)
colnames(obj.THCA.nkt@meta.data) <- colnames(obj.THCA.nkt@meta.data) %>%␣
 ↪gsub("integrated_snn_res.0.", "nkt0", .)
```

[ ]:
```
obj.THCA.nkt@meta.data <- obj.THCA.nkt@meta.data %>%
                          mutate(dCT = case_when(nkt04 %in% c('5', '8') ~ 'NK',
                                                 nkt04 %in% c('3', '6') ~␣
 ↪'Treg',
                                                 nkt04 %in% c('1', '2', '7') ~␣
 ↪'CD4T',
                                                 nkt04 %in% c('0', '4') ~␣
 ↪'CD8T',
                                                 TRUE ~ 'Others'
                                                 ))
obj.THCA.nkt@meta.data <- obj.THCA.nkt@meta.data %>%
                          dplyr::select(-c('nkte01', 'nkte02', 'nkte03',␣
 ↪'nkte04', 'seurat_clusters', 'r02', 'nkt01', 'nkt02', 'nkt03', 'nkt04'))
```

3.3 assign Myeloid/B/Plasma

[ ]:
```
marker_list <- c('PTPRC', 'LYZ',
                 'CD163','CD68','ITGAX','MARCO', 'MRC1', 'SLC40A1', 'SPP1', ##␣
 ↪Mph
                 'S100A8', 'S100A9', 'THBS1', 'CD14', 'FCGR3A','FCN1', 'VCAN',␣
 ↪## Myeloid
                 'CD1C', 'IDO1', 'CLEC4C', 'CSF2RA', 'LAMP3', 'CLEC10A', ## DC
                 'CD79A', 'CD79B', 'MS4A1', 'IGHM', 'IGHD', 'IGKC', 'CD74',␣
 ↪'JCHAIN')
```

[ ]:
```
obj.THCA.mb <- obj.THCA %>% subset(mCT %in% c('Myeloid', 'B/Plasma'))
DefaultAssay(obj.THCA.mb) <- 'integrated'
# scale
obj.THCA.mb <- obj.THCA.mb %>%
                 FindVariableFeatures(selection.method = "vst", nfeatures =␣
 ↪1000, verbose = F) %>%
                 ScaleData(verbose = FALSE) %>%
                 RunPCA(npcs = 50, verbose = F)
# PCA+clustering
select <- 1:(PC_selection(obj.THCA.mb)$PCselect %>% min())
obj.THCA.mb <- obj.THCA.mb %>%
                 RunUMAP(reduction = "pca", dims = select, umap.method =␣
 ↪"uwot") %>%
                 RunTSNE(reduction = "pca", dims = select) %>%
                 FindNeighbors(reduction = "pca", dims = select) %>%
                 FindClusters(resolution = 0.1) %>% FindClusters(resolution = 0.
 ↪2) %>% FindClusters(resolution = 0.3) %>% FindClusters(resolution = 0.4)
colnames(obj.THCA.mb@meta.data) <- colnames(obj.THCA.mb@meta.data) %>%␣
 ↪gsub("integrated_snn_res.0.", "mb0", .)
```

[ ]:
```
Idents(obj.THCA.mb) <- factor(obj.THCA.mb$mb02, levels = 8:0)
cluster_deg <- FindAllMarkers(obj.THCA.mb, assay = "RNA", slot = "data",
                              logfc.threshold = 0.25, min.pct = 0.1, test.use =␣
 ↪"wilcox")
rownames(cluster_deg) <- NULL
```

[ ]:
```
obj.THCA.mb@meta.data <-  obj.THCA.mb@meta.data %>%
                          mutate(dCT = case_when(mb02 %in% c('0', '2', '8') ~␣
 ↪'B',
                                                 mb02 %in% c('1', '6') ~␣
 ↪'Plasma',
                                                 mb02 %in% c('3') ~ 'Mph',
                                                 mb02 %in% c('4') ~ 'Mono',
                                                 mb02 %in% c('5', '7') ~ 'DC',
                                                 TRUE ~ 'Others'
                                                 ))
```

```
obj.THCA.mb@meta.data <- obj.THCA.mb@meta.data %>% dplyr::select(-c('mb01',␣
↪'mb02', 'mb03', 'mb04', 'seurat_clusters', 'r02'))
```

3.4 assign oCT/gCT

[ ]:
```
obj.THCA.others <- subset(obj.THCA, barcode %in% c(obj.THCA.epi$barcode, obj.
↪THCA.nkt$barcode, obj.THCA.mb$barcode) == FALSE)
obj.THCA.others@meta.data <- obj.THCA.others@meta.data %>%
                             dplyr::select(-c('r02')) %>%
                             transform(dCT = mCT)
```

[ ]:
```
meta_new <- list(obj.THCA.epi@meta.data, obj.THCA.nkt@meta.data, obj.THCA.
↪mb@meta.data, obj.THCA.others@meta.data) %>%
            do.call(rbind, .) %>%
            .[colnames(obj.THCA),]
obj.THCA@meta.data <- meta_new
obj.THCA@meta.data <- obj.THCA@meta.data %>%
                      transform(oCT = dCT) %>%
                      mutate(mCT = case_when(dCT %in% c('B', 'Plasma') ~ 'B',
                                             TRUE ~ dCT)) %>%
                      mutate(gCT = case_when(mCT %in% c('Epi') ~ 'Tumor',
                                             mCT %in% c('CD4T', 'Treg', 'B',␣
↪'Plasma', 'CD8T', 'Mph', 'DC', 'NK', 'Mono') ~ 'Immune',
                                             mCT %in% c('Endo', 'Fibro') ~␣
↪'Stromal',
                                             TRUE ~ 'Others'))
```

4 save

[ ]:
```
saveRDS(obj.THCA, 'obj.THCA.use.rds')
```