# LIHC_Xue2022_process

December 25, 2025

## 1 sample match

[ ]:
```r
# According to the supp1 table of article+meta info of GSA
meta1 <- read.xlsx('/project/sex_cancer/data/LIHC_Xue2022/supp1_patient_info.xlsx', sheet = 2,startRow = 3) %>%
        subset(Cancer_type == 'Hepatocellular carcinoma') %>% subset(M == 0)
meta2 <- read.xlsx('/project/sex_cancer/data/LIHC_Xue2022/supp1_patient_info.xlsx', sheet = 3,startRow = 3, rowNames = T) %>%
        transform(Patient = strsplit2(Sample, split = '_')[,1]) %>%␣
↪subset(Patient %in% meta1$Patient) %>%
        dplyr::select(c('Patient', 'Sample', 'Number.of.cells.after.QC',␣
↪'Number.of.Reads', 'Total.Genes.Detected'))
meta3 <- merge(meta2, meta1, by = 'Patient', all = TRUE)
meta4 <- read.xlsx('/project/sex_cancer/data/LIHC_Xue2022/GSA_HRA001748.xlsx',␣
↪sheet = 4) %>%
        dplyr::select(c('Accession', 'Run.title')) %>% dplyr::
↪rename(c('Sample' = 'Run.title')) %>%
        subset(Sample %in% meta3$Sample)

meta <- merge(meta4, meta3, by = 'Sample') %>% .[grepl('_HCC$', .$Sample),]
select.F <- meta %>% subset(Gender == 'F') %>% .[order(.$Accession),] %>% .
↪[order(-.$Number.of.cells.after.QC),] %>% subset(Number.of.cells.after.
↪QC>5000)
select.M <- meta %>% subset(Gender == 'M') %>% .[order(.$Accession),] %>% .
↪[order(-.$Number.of.cells.after.QC),] %>% subset(Number.of.cells.after.
↪QC>5000)
match_data <- rbind(select.F, select.M) %>%
              transform(Sex = ifelse(Gender == 'F', 1, 0)) %>%
              dplyr::rename(c('Age' = "Age..years."))
head(match_data, n = 2)

match_model <- matchit(formula = Sex ~ Age,
                       data = match_data,
                       method = "nearest",
                       distance = "logit",
                       ratio = 1,
                       replace = FALSE)
```

```
matched_data <- match_data(match_model)

matched_data <- matched_data %>% rename_with(~ str_replace_all(.x, "\\.{1,}",␣
↪"_") |> str_replace("_$", ""))
```

## 2 load data

[ ]:
```
obj.LIHC <- readRDS('/project/sex_cancer/data/LIHC_Xue2022/Seurat_com.counts.
↪seurat_rm.hvg2000_PC15_res1.rds')
obj.LIHC
```

## 3 modify meta.data

[ ]:
```
meta <- matched_data
## select partial info
meta_select <- meta %>% dplyr::select('Sample', 'Accession', 'Gender', 'Age',␣
↪'Tumor_size_cm', 'Differentiation',
                                      'Venous_invasion', 'T', 'N', 'M',␣
↪'TNM_stage', 'BCLC_stage',
                                      'Cirrhosis', 'Virus',␣
↪'Relapse_state_Yes_0', 'FPS_time', 'OS_state_Yes_0', 'OS_time')

## extract barcode order
cell_order <- rownames(obj.LIHC@meta.data)
obj.LIHC@meta.data <- obj.LIHC@meta.data %>%
                      transform(barcode = rownames(.), Accession = orig.ident)␣
↪%>%
                      merge(., meta_select, by = 'Accession', all = TRUE) %>%
                      dplyr::select(-'Sample.x') %>% dplyr::rename(c('Sample' =␣
↪'Sample.y')) %>%
                      `rownames<-`(., .$barcode) %>%
                      dplyr::rename(c('Sex' = 'Gender', 'SampleID' = 'orig.
↪ident')) %>%
                      transform(Cohort = "LIHC_Xue2022") %>%
                      transform(SampleType = 'Tumor')  %>%
                      transform(Chemistry = "10x 3' v3") %>%
                      .[cell_order,]
```

## 4 cell type annotation

[ ]:
```
obj.LIHC <- obj.LIHC %>%
            NormalizeData(normalization.method = "LogNormalize", scale.factor =␣
↪10000, verbose = F) %>%
```

```
            FindVariableFeatures(selection.method = "vst", nfeatures = 2000,␣
 ↪verbose = F) %>%
            ScaleData(vars.to.regress = c("percent.mt"), verbose = F) %>%
            RunPCA(verbose = F) %>%
            RunHarmony(group.by.vars = "Accession", plot_convergence = TRUE)

## cluster
nPC <- min(PC_selection_harmony(obj.LIHC)$PCselect)
obj.LIHC <- obj.LIHC %>%
            RunUMAP(reduction = "harmony", dims = 1:nPC, umap.method = "uwot")␣
 ↪%>%
            RunTSNE(reduction = "harmony", dims = 1:nPC) %>%
            FindNeighbors(reduction = "harmony", dims = 1:nPC) %>%
            FindClusters(resolution=0.1) %>% FindClusters(resolution=0.2) %>%␣
 ↪FindClusters(resolution=0.3)
colnames(obj.LIHC@meta.data) <- gsub("RNA_snn_res.0.","r0", colnames(obj.
 ↪LIHC@meta.data))
```

[ ]:
```
Idents(obj.LIHC) <- factor(obj.LIHC$r01, levels = 15:0)
cluster_deg <- FindAllMarkers(obj.LIHC, assay = "RNA", slot = "data",
                              logfc.threshold = 0.25, min.pct = 0.1, test.use =␣
 ↪"wilcox")
rownames(cluster_deg) <- NULL
```

[ ]:
```
## marker expression
marker_list <- c('PTPRC', 'CD3D', 'CD8A', 'CD4', 'NKG7', 'TRDC', 'FOXP3',␣
 ↪'CD79A', 'MS4A1', ## lymphoid
                 'CD14', 'CD68', 'CD163', 'CD1C', 'LAMP3', 'CSF3R', 'S100A8',␣
 ↪'CLEC10A', 'TPSAB1', ## myeloid 'CD16', 'TPSAB1',
                 'VWF', 'COL1A1', # stromal cell
                 'ALB', 'EPCAM') ## epithelial
```

[ ]:
```
## drop C15
obj.LIHC <- obj.LIHC %>% subset(r01 %in% c('15') == FALSE)
obj.LIHC@meta.data <- obj.LIHC@meta.data %>%
                      mutate(mCT = case_when(r01 %in% c('2', '10', '14') ~␣
 ↪'Epi',
                                             r01 %in% c('0', '4', '5', '8') ~␣
 ↪'NK/T',
                                             r01 %in% c('1', '11', '12', '13')␣
 ↪~ 'Myeloid',
                                             r01 %in% c('7', '9') ~ 'B',
                                             r01 %in% c('3') ~ 'Endothelial',
                                             r01 %in% c('6') ~ 'Fibroblast',
                                             # r01 %in% c('15') ~ 'Others',
                                             TRUE ~ 'Others'
```

```
                                                )) %>%
                        transform(mCT =  factor(mCT, levels = c('Tumor', 'NK/
 ↪T', 'B', 'Myeloid', 'Endothelial', 'Fibroblast')))

obj.LIHC@meta.data <- obj.LIHC@meta.data %>%
                      dplyr::select(-c('r03', 'r02', 'RNA_snn_res.1',␣
 ↪'seurat_clusters', 'Accession'))
```

### 4.1 assign Epi

[ ]:
```
obj.LIHC.epi <- obj.LIHC %>% subset(r01 %in% c(2,10,14))
obj.LIHC.epi
```

### 4.2 assign Myeloid

[ ]:
```
marker_list <- c('CD163','CD68','ITGAX','MARCO', 'MRC1', 'SLC40A1', 'SPP1', ##␣
 ↪Mph
                 'CD14', 'FCGR3A','FCN1', 'VCAN',  ## monocyte
                 'TPSAB1', ## mast cell
                 'CSF3R', 'S100A8', 'S100A9', ## neutrophil
                 'CD1C', 'IDO1', 'CLEC4C', 'CSF2RA', 'LAMP3', 'CLEC10A')  ## DC
```

[ ]:
```
obj.LIHC.mye <- obj.LIHC %>% subset(mCT == 'Myeloid')
obj.LIHC.mye

## anchor-based integration
sampleList = unique(ext_list(obj.LIHC.mye$orig.ident)); sampleList;␣
 ↪length(sampleList)
obj.anchor <- lapply(sampleList, function(sampleID){
                      obj <- obj.LIHC.mye %>%
                             subset(orig.ident == sampleID) %>%
                             NormalizeData(normalization.method =␣
 ↪"LogNormalize", scale.factor = 10000, verbose = F) %>%
                             FindVariableFeatures(selection.method = "vst",␣
 ↪nfeatures = 1000, 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
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.", "myeR0", .)
```

[ ]:
```
options(repr.plot.height = 6, repr.plot.width = 8)
VlnPlot(obj.anchor, group.by = 'myeR03', features = marker_list, pt.size = 0,␣
 ↪raster=FALSE, stack = TRUE, flip = TRUE)+theme(legend.position =  'none')
```

[ ]:
```
Idents(obj.anchor) <- factor(obj.anchor$myeR03, levels = 14:0)
cluster_deg <- FindAllMarkers(obj.anchor, assay = "RNA", slot = "data",
                              logfc.threshold = 0.25, min.pct = 0.1, test.use =␣
 ↪"wilcox")
rownames(cluster_deg) <- NULL
```

[ ]:
```
## assign annotation
obj.LIHC.mye <- obj.anchor
obj.LIHC.mye@meta.data <- obj.LIHC.mye@meta.data %>%
                          mutate(oCT = case_when(myeR03 %in% c('0', '1', '4',␣
 ↪'6', '9') ~ 'Mph',
                                                 myeR03 %in% c('2', '7') ~␣
 ↪'Mono',
                                                 myeR03 %in% c('13') ~ 'Mast',
                                                 myeR03 %in% c('5', '10') ~␣
 ↪'Neu',
                                                 myeR03 %in% c('3', '8', '11',␣
 ↪'12', '14') ~ 'DC'
                                                 ))
table(obj.LIHC.mye$myeR03, obj.LIHC.mye$oCT)
obj.LIHC.mye@meta.data <- obj.LIHC.mye@meta.data %>% dplyr::select(-c('myeR02',␣
 ↪'seurat_clusters', 'myeR01', 'myeR04'))
```

### 4.3 assign NK/T

[ ]:
```
marker_list <- c('CD3D', 'CD3E', 'CD3G', 'TRDC', ## T cell
                 'CD4','FOXP3', 'CTLA4', ## Treg: 'FOXP3', 'CTLA4'
                 'CD8A', 'CD8B', 'CD28', 'GZMA', 'GZMH', ## CD8T
                 'TIGIT', 'PDCD1', 'TCF7',
                 'GNLY', 'NKG7', 'KLRD1', 'NCAM1', 'FCGR3A', 'PRF1', ## NK␣
 ↪NCAM1-CD56 FCGR3A-CD16
                 'MKI67', 'TOP2A', 'STMN1', 'TOX')
```

[ ]:
```
obj.LIHC.nkt <- obj.LIHC %>% subset(mCT == 'NK/T')
sampleList = unique(ext_list(obj.LIHC.nkt$orig.ident)); sampleList;␣
 ↪length(sampleList)
obj.anchor <- lapply(sampleList, function(sampleID){
                       obj <- obj.LIHC.nkt %>%
                              subset(orig.ident == sampleID) %>%
                              NormalizeData(normalization.method =␣
 ↪"LogNormalize", scale.factor = 10000, verbose = F) %>%
                              FindVariableFeatures(selection.method = "vst",␣
 ↪nfeatures = 1000, 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
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.", "nktR0", .)
```

[ ]:
```
Idents(obj.anchor) <- factor(obj.anchor$nktR03, levels = 12:0)
cluster_deg <- FindAllMarkers(obj.anchor, assay = "RNA", slot = "data",
                              logfc.threshold = 0.25, min.pct = 0.1, test.use =␣
 ↪"wilcox")
rownames(cluster_deg) <- NULL
```

[ ]:
```
obj.LIHC.nkt <- obj.anchor
obj.LIHC.nkt@meta.data <- obj.LIHC.nkt@meta.data %>%
                          mutate(oCT = case_when(nktR03 %in% c('0', '10') ~␣
 ↪'CD4T',
                                                 nktR03 %in% c('1', '4', '5',␣
 ↪'8', '9', '11') ~ 'CD8T',
                                                 nktR03 %in% c('2') ~ 'Treg',
                                                 nktR03 %in% c('3') ~ 'NK',
                                                 nktR03 %in% c('7') ~ 'MAIT',
                                                 nktR03 %in% c('6', '12') ~␣
 ↪' T',
                                                 ))
obj.LIHC.nkt@meta.data <- obj.LIHC.nkt@meta.data %>%
                          dplyr::select(-c('seurat_clusters', 'nktR01',␣
 ↪'nktR02', 'nktR04'))
```

### 4.4 assign mCT/gCT

[ ]:
```
obj.LIHC.other <- obj.LIHC %>%
                  subset(barcode %in% c(obj.LIHC.nkt$barcode, obj.LIHC.
 ↪mye$barcode) == FALSE)
```

[ ]:
```
obj.LIHC.other@meta.data <- obj.LIHC.other@meta.data %>%
                            mutate(oCT = case_when(r01 ==  '9' ~ 'plasma',
                                                   TRUE ~ mCT)) %>%
                            dplyr::select(-'r01')
obj.LIHC.nkt@meta.data <- obj.LIHC.nkt@meta.data %>% dplyr::select(-c('r01',␣
 ↪'nktR03'))
obj.LIHC.mye@meta.data <- obj.LIHC.mye@meta.data %>% dplyr::select(-c('r01',␣
 ↪'myeR03'))

obj.LIHC <- merge(obj.LIHC.other, c(obj.LIHC.nkt, obj.LIHC.mye))
```

[ ]:
```
obj.LIHC@meta.data  <-  obj.LIHC@meta.data %>%
                            mutate(mCT = case_when(oCT ==  'Endothelial' ~␣
 ↪'Endo',
                                                   oCT ==  'Fibroblast' ~␣
 ↪'Fibro',
                                                   oCT ==  'plasma' ~ 'B',
                                                   TRUE ~ oCT
```

```
                                                    ))
head(obj.LIHC@meta.data, n  = 2)
```

[ ]:
```
obj.LIHC@meta.data  <-  obj.LIHC@meta.data %>%
                            mutate(gCT = case_when(mCT %in% 'Epi' ~ 'Tumor',
                                                   mCT %in% c('B', 'NK',␣
 ↪'CD8T', 'Treg', 'CD4T', ' T', 'MAIT', 'DC', 'Mph', 'Mono', 'Neu', 'Mast') ~␣
 ↪'Immune',
                                                   mCT %in% c('Endo', 'Fibro')␣
 ↪~ 'Stromal',
                                                   TRUE ~ 'Others'))
head(obj.LIHC@meta.data, n  = 2)
```

## 5 UMAP visualization

[ ]:
```
mat <- obj.LIHC@reductions$pca@cell.embeddings[colnames(obj.LIHC),]
obj.LIHC[['pca']] <- CreateDimReducObject(embeddings = mat, key = 'PC_', assay␣
 ↪= 'RNA')

mat <- obj.LIHC@reductions$umap@cell.embeddings[colnames(obj.LIHC),]
obj.LIHC[['umap']] <- CreateDimReducObject(embeddings = mat, key = 'umap_',␣
 ↪assay = 'RNA')

mat <- obj.LIHC@reductions$tsne@cell.embeddings[colnames(obj.LIHC),]
obj.LIHC[['tsne']] <- CreateDimReducObject(embeddings = mat, key = 'tSNE_',␣
 ↪assay = 'RNA')
```

[ ]:
```
options(repr.plot.height = 5, repr.plot.width = 25)
DimPlot_scCustom(obj.LIHC, pt.size = 1, group.by = "oCT", reduction = 'umap',␣
 ↪label = T, label.size = 4, colors_use = pal_igv("default")(51))|
DimPlot_scCustom(obj.LIHC, pt.size = 1, group.by = "mCT", reduction = 'umap',␣
 ↪label = T, label.size = 4, colors_use = pal_igv("default")(51))|
DimPlot_scCustom(obj.LIHC, pt.size = 1, group.by = "gCT", reduction = 'umap',␣
 ↪label = T, label.size = 4, colors_use = pal_igv("default")(51))
```

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