Quality Control for scRNA and scATAC seq data
Merge the data

Ximing Ran

2025-05-01

Contents

```r
library(Seurat)
```

```
## Warning: package 'Seurat' was built under R version 4.4.1

## Loading required package: SeuratObject

## Loading required package: sp

## Warning: package 'sp' was built under R version 4.4.1

## 'SeuratObject' was built with package 'Matrix' 1.7.0 but the current
## version is 1.7.3; it is recomended that you reinstall 'SeuratObject' as
## the ABI for 'Matrix' may have changed

##
## Attaching package: 'SeuratObject'

## The following objects are masked from 'package:base':
##
##     intersect, t
```

```r
library(tidyverse)
```

```r
# load function from local files
```

```r
# load data from local files
library(Signac)
```

```
## Warning: package 'Signac' was built under R version 4.4.1
```

```r
library(Seurat)
```

1. Read the QC Data

```r
options(Seurat.object.assay.version = "v5")

# load data
sample_seurat_list <- readRDS(
  here::here("data", "processed-data",
             "sample_seurat_RNA_list_2025-05-06.rds")
)

sample_meta <- readRDS(
  here::here("data", "processed-data",
             "sample_meta_RNA_2025-05-06.rds")
)




for (samp in names(sample_seurat_list)) {
  cat("Sample:", samp, "\n")
  print(head(colnames(sample_seurat_list[[samp]]), 10))
  cat("\n")
}
```

```
## Sample: CTRL
##  [1] "CTRL_AAACAGCCATCAGTAT-1" "CTRL_AAACAGCCATTGGGAG-1"
##  [3] "CTRL_AAACATGCAAGCCAGA-1" "CTRL_AAACATGCACAATACT-1"
##  [5] "CTRL_AAACCAACAATTATGC-1" "CTRL_AAACCAACACTATGGC-1"
##  [7] "CTRL_AAACCAACAGACAAAC-1" "CTRL_AAACCAACAGCAACAG-1"
##  [9] "CTRL_AAACCGAAGACCATAC-1" "CTRL_AAACCGAAGCTATTGA-1"
##
## Sample: TSC_edge
##  [1] "TSC_edge_AAACAGCCACCATATG-1" "TSC_edge_AAACAGCCACGAATCC-1"
##  [3] "TSC_edge_AAACAGCCAGCTCATA-1" "TSC_edge_AAACAGCCAGGCTTCG-1"
##  [5] "TSC_edge_AAACAGCCATAATTGC-1" "TSC_edge_AAACAGCCATCAGCAC-1"
##  [7] "TSC_edge_AAACAGCCATGATTGT-1" "TSC_edge_AAACAGCCATTGTTGG-1"
##  [9] "TSC_edge_AAACATGCAAAGCGGC-1" "TSC_edge_AAACATGCACAATGCC-1"
##
## Sample: TSC_tuber
##  [1] "TSC_tuber_AAACAGCCAAATGCCC-1" "TSC_tuber_AAACAGCCAACCTGGT-1"
##  [3] "TSC_tuber_AAACAGCCAAGACTCC-1" "TSC_tuber_AAACAGCCAAGCTTTG-1"
##  [5] "TSC_tuber_AAACAGCCAATTAGCT-1" "TSC_tuber_AAACAGCCATAAACCT-1"
##  [7] "TSC_tuber_AAACAGCCATAACGGG-1" "TSC_tuber_AAACAGCCATCCCTCA-1"
##  [9] "TSC_tuber_AAACATGCAGCATGGA-1" "TSC_tuber_AAACATGCATTATGCG-1"
##
## Sample: TSC_outside
##  [1] "TSC_outside_AAACAGCCAAATGCCC-1" "TSC_outside_AAACAGCCAACCTGGT-1"
##  [3] "TSC_outside_AAACAGCCAAGACTCC-1" "TSC_outside_AAACAGCCAAGCTTTG-1"
##  [5] "TSC_outside_AAACAGCCAATTAGCT-1" "TSC_outside_AAACAGCCAGCCTAAC-1"
##  [7] "TSC_outside_AAACAGCCATAAACCT-1" "TSC_outside_AAACAGCCATAACGGG-1"
##  [9] "TSC_outside_AAACAGCCATCCCTCA-1" "TSC_outside_AAACATGCACAAAGGT-1"
```

2. Merge the data

```r
library(Seurat)
library(harmony)
```

```
## Warning: package 'harmony' was built under R version 4.4.1

## Loading required package: Rcpp

## Warning: package 'Rcpp' was built under R version 4.4.1
```

```r
library(Signac)
library(GenomicRanges)
```

```
## Warning: package 'GenomicRanges' was built under R version 4.4.1

## Loading required package: stats4

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following object is masked from 'package:SeuratObject':
##
##     intersect

## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union

## The following objects are masked from 'package:dplyr':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, table,
##     tapply, union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

## Warning: package 'S4Vectors' was built under R version 4.4.1

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:lubridate':
##
##     second, second<-

## The following objects are masked from 'package:dplyr':
##
##     first, rename
```

```
## The following object is masked from 'package:tidyr':
##
##     expand

## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

## Warning: package 'IRanges' was built under R version 4.4.1

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:sp':
##
##     %over%

## The following object is masked from 'package:lubridate':
##
##     %within%

## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice

## The following object is masked from 'package:purrr':
##
##     reduce

## Loading required package: GenomeInfoDb
```

```r
library(dplyr)

names(sample_seurat_list) <- sample_meta

# 1) Pre-process each sample's RNA (normalize -> HVG -> scale -> PCA)

sample_seurat_list <- lapply(sample_seurat_list, function(obj) {
  DefaultAssay(obj) <- "RNA"
  obj %>%
    NormalizeData(assay = "RNA", layer = "counts.Gene Expression") %>%
    FindVariableFeatures(assay = "RNA", layer = "counts.Gene Expression",
                         nfeatures = 3000) %>%
    ScaleData(assay = "RNA", layer = "counts.Gene Expression") %>%
    RunPCA(assay = "RNA", reduction.name = "pca", npcs = 50)
})
```

```
## Normalizing layer: counts.Gene Expression

## Finding variable features for layer counts.Gene Expression

## Centering and scaling data matrix

## PC_ 1
## Positive:  ST18, DOCK5, CNDP1, SLC5A11, FOLH1, MOBP, LINC01608, FAM107B, RNF220, AL033523.1
```

```
##     CLDN11, TMEM144, ACSBG1, DOCK8, ITGA2, C3, HHIP, APBB1IP, CSF1R, RGS1
##     CSF2RA, SYK, RHBDF2, FYB1, CX3CR1, AC079352.1, CPM, LINC00609, PIK3R5, LINC02642
## Negative: DGKI, LDB2, NEK10, CHRM3, GABRB2, PAK5, PPM1L, CACNA1E, CHN1, SYBU
##     LRRTM4, AL008633.1, SORBS2, EPB41L3, PTPRR, KCNJ3, KCNQ5, KCTD16, LINGO2, RGS6
##     EPB41L4B, KIAA1211L, ZNF385B, SH3GL2, MLIP, VSNL1, ADGRL2, CNTN4, NELL2, NELL1
## PC_ 2
## Positive: AL117329.1, DLGAP1-AS4, NWD2, KIAA1211L, AC011995.2, LINC00507, MCHR2, MLIP, AC073091.3, L
##     CUX2, NGEF, ENC1, SLC22A10, AC011287.1, CDH22, CABP1, AC073365.1, LAMB1, GSG1L
##     TAFA1, RFTN1, ARHGAP10, HRH2, SATB2, TESPA1, IQCJ-SCHIP1, LINC01250, AP000829.1, FSTL4
## Negative: PREX2, GPC5, SLC1A2, NKAIN3, PARD3B, LIFR, ATP13A4, SLC1A3, GLI3, NHSL1
##     HIF3A, ADGRV1, PARD3, ATP1A2, NEAT1, ACACB, STON2, AL137139.2, SLCO1C1, LINC00299
##     OBI1-AS1, CRB1, HPSE2, ACSS1, RANBP3L, RFX4, PLPP3, GJA1, ARHGEF26, LINC00472
## PC_ 3
## Positive: SNRPN, GAD1, GRIP1, ZNF536, ANK1, RAB3IP, VAV2, ZNF385D, NXPH1, KIAA1211
##     GRIN2D, VWC2, GRIP2, KLHL5, NHS, SILC1, DNER, LRP8, UBASH3B, GAD2
##     SPOCK3, DOCK11, PTCHD4, PTPRM, BCL11B, GRIK3, CFAP161, ELAVL2, SLC8A3, PCP4L1
## Negative: BMPR1B, SHROOM3, ADGRV1, SERPINE2, SLC1A2, MEIS2, RGS20, ZNRF3, SLC1A3, LINC00299
##     NHSL1, GPC5, LINC00507, ATP13A4, GLIS3, COL5A3, TAFA1, AL589740.1, STON2, GLI3
##     ATP1A2, ACSS1, SLC4A4, SLC25A18, OBI1-AS1, SLCO1C1, ARHGEF26, PAMR1, RANBP3L, CDH20
## PC_ 4
## Positive: SLCO2B1, INPP5D, SRGN, LRMDA, ST6GAL1, DOCK8, FLI1, PALD1, DLEU1, IL13RA1
##     IPCEF1, APBB1IP, ADAM28, P2RY12, IRAK3, A2M, TGFBR2, TBXAS1, SP100, DOCK2
##     C3, AOAH, FYB1, LINC02712, SYK, PTPRC, CSF1R, RHBDF2, AC012150.1, CSF2RA
## Negative: NKAIN3, ATP13A4, SLC1A2, ACSBG1, CRB1, LRP4, ADGRV1, COL5A3, CDH20, RFX4
##     LINC00299, SLC25A18, GPC5, AC079352.1, HPSE2, SLCO1C1, ARHGEF26, AC091826.2, ETNPPL, ALDH1L1
##     LINC00499, PLPP3, FREM2, PAMR1, OBI1-AS1, SLC25A48, PTPRZ1, C10orf90, SPON1, ARHGEF26-AS1
## PC_ 5
## Positive: RASGRF2, DAB1, HS6ST3, SLIT2, GRIA4, KIF26B, EPHA6, LINC01331, KCNMB2, ADAMTS9-AS2
##     GAD1, ZNF536, PDZD2, FAM163A, PDGFD, CDH13, CNTN6, THSD7A, LINC00507, GRIP2
##     PCSK6, ZNF804A, NECTIN3, ROBO1, NDST3, TTC6, SLC24A4, GALNTL6, PRKG2, LAMB1
## Negative: FOXP2, CLSTN2, PDZRN4, PTK7, SLIT1, HS3ST2, RXFP1, HS3ST4, IL1RAPL2, AL353709.1
##     RORB, AC008415.1, CASC15, LINC02232, KHDRBS3, POU6F2, AC109466.1, TSHZ2, CPNE4, TRMT9B
##     NPTX1, AL033539.2, RALYL, NELL2, ATP8B1, AC073578.2, TLL1, LRRK1, PDE1A, SLC22A10

## Normalizing layer: counts.Gene Expression

## Finding variable features for layer counts.Gene Expression

## Centering and scaling data matrix

## PC_ 1
## Positive: AC008571.2, LINC01170, PIK3R5, DOCK8, PTPRC, IKZF1, LINC01374, APBB1IP, LINC01445, TBXAS1
##     LNCAROD, P2RY12, VSIG4, SYK, PIK3AP1, ADAM28, OLR1, SPP1, TLR2, TFEC
##     BCAS1, TMEM156, FYB1, RGS1, LYN, LINC02712, AC079015.1, AGMO, SCIN, AC012150.1
## Negative: SNHG14, MYT1L, SCN2A, ATP8A2, SLC25A12, PTPRN2, MAP2, NBEA, UNC80, SNAP91
##     GRIN2B, CACNA1C, AGBL4, NRG3, SCN3A, STXBP1, SPTBN4, NRXN1, RIMS2, MAPK10
##     PAK3, CACNA1B, FAM155A, KCNMA1, CDK14, PCLO, ADGRL3, TMEM178B, DLGAP1, PLCB1
## PC_ 2
## Positive: SORBS1, DTNA, NPL, GPM6A, RFX4, SLC14A1, TPD52L1, PITPNC1, ACSS3, RYR3
##     BMPR1B, GLIS3, AL137139.2, ABLIM1, NEBL, AC012405.1, PRKG1, AQP4, STON2, PLCE1
##     FMN2, CD44, DOCK7, ARHGEF4, WDR49, SHROOM3, SOX5, ADCY2, CTNNA2, AC073941.1
## Negative: LRRC7, KCNQ3, HDAC9, PLCL2, IGSF21, GRID2, TMEFF2, MDGA2, MYT1L, NRXN3
##     LINC00632, SRGAP2, FRRS1L, KHDRBS3, SCN3A, MEF2C, ATP8A2, DGKI, GRIN2B, CACNA1D
##     NECAB1, SLC8A1, RAB11FIP4, SRRM3, RIMS2, SYT16, PTPRD, NLGN1, LHFPL4, PTPRN2
## PC_ 3
## Positive: GALNT13, NLGN1, SOX2-OT, SMOC1, LINC01170, LHFPL3, PCDH15, TNR, LRRTM3, DCC
```

```
##     BCAS1, LUZP2, TMEFF2, MEGF11, ERBB4, AL392023.2, MDGA2, MYT1, CADM2, AC008571.2
##     KCNMB2, NLGN4X, SEMA5A, PDGFRA, CA10, CSMD2, XYLT1, CACNG4, SGCD, VCAN
## Negative:  LRMDA, SRGAP2, PLXDC2, SRGAP2B, DOCK8, ARHGAP24, SMAP2, ATP8B4, APBB1IP, LPCAT2
##     ST6GAL1, IRAK3, PALD1, PRKCH, LINC01374, TBXAS1, LYN, ARHGAP15, LNCAROD, SKAP2
##     ADAM28, PDE3B, TGFBR2, PIK3R5, LHFPL2, OLR1, ABCC4, P2RY12, RGS1, IPCEF1
## PC_ 4
## Positive:  PDGFRA, PCDH15, LHFPL3, TMEM132C, TNR, MEGF11, PDZD2, AFAP1L2, LUZP2, PTPRG
##     ZEB1, PLCB4, ATP10A, RBMS3, COBLL1, MYT1, CRISPLD2, TEK, PTPRM, FAM110B
##     EPAS1, ANO2, BX284613.2, CACNG4, LAMA4, ARHGAP29, SLC35F1, XYLT1, COL9A1, VCAN
## Negative:  PTPRD, NECAB1, SATB2, PPP2R2C, SLC17A7, DPP10, NELL2, AC011995.2, SH3GL2, AGBL4
##     ADAMTS16, NPTX1, SV2B, NGEF, RALYL, AC067956.1, CAMKK1, SYT7, PHACTR1, CCBE1
##     LINC02217, CAMK2A, RYR2, SLIT3, CELF2, LY86-AS1, CCDC3, NRGN, B4GALNT3, SLC8A2
## PC_ 5
## Positive:  ATP10A, TACC1, ABCB1, EPAS1, COBLL1, CLDN5, PTPRB, ITGA1, MECOM, CADPS2
##     ARHGAP29, THSD4, ADGRF5, ABCG2, PLEKHG1, ELOVL7, NOSTRIN, PALMD, LEF1, CLEC1A
##     EBF1, ITIH5, PODXL, BTNL9, VWF, ESYT2, PLCB4, EPHA4, ANXA3, CFH
## Negative:  PCDH15, LHFPL3, TNR, MEGF11, LUZP2, PTPRZ1, VCAN, MYT1, DSCAM, SLC35F1
##     FAM110B, XYLT1, KCNMB2, CA10, CACNG4, OPHN1, PLAAT1, AC004852.2, COL11A1, PDGFRA
##     COL9A1, BX284613.2, CHST9, GRID2, SMOC1, SOX6, FERMT1, AL512308.1, LRRTM3, STK32A

## Normalizing layer: counts.Gene Expression

## Finding variable features for layer counts.Gene Expression

## Centering and scaling data matrix

## PC_ 1
## Positive:  AC008571.2, SPP1, BCAS1, APBB1IP, DOCK8, RGS1, LNCAROD, PDK4, AP000477.3, PIK3R5
##     LINC01445, OLR1, TGFBR2, PTPRC, SCIN, LINC02642, TMEM156, AC012150.1, AC079015.1, AL392172.2
##     FYB1, LPCAT2, VSIG4, CCL3, ITGAX, LINC01374, CCL4L2, ADAM28, PIK3AP1, TBXAS1
## Negative:  SCN2A, SNHG14, PTPRN2, ATP8A2, GRIN2B, MYT1L, FAM155A, LRRC7, SLC4A10, CACNA1C
##     UNC80, KSR2, UNC79, RBFOX1, GABBR2, ANKS1B, CSMD1, DLGAP2, SCN8A, GABRB3
##     MAP2, CACNA1B, RIMS2, NBEA, KCNMA1, RGS7, TTC3, C2CD5, MSRA, KALRN
## PC_ 2
## Positive:  RFX4, RYR3, AL137139.2, ACSS3, TPD52L1, GLIS3, GPM6A, SLC14A1, AC012405.1, AQP4
##     PRKG1, WDR49, MRVI1, CD44, AC073941.1, PLSCR4, BMPR1B, LINC00299, AC091078.1, CTNNA2
##     SLC16A9, COLEC12, AC053513.1, AC002429.2, ZNRF3, RNF19A, ETNPPL, SHROOM3, FAM189A2, FAM107A
## Negative:  SMAP2, SRGAP2B, LPCAT2, DOCK8, PDE3B, ATP8B4, LINC01374, RGS1, IPCEF1, APBB1IP
##     PTPRC, ADAM28, ARHGAP15, LNCAROD, LRMDA, PALD1, HDAC9, PIK3R5, LYN, IKZF1
##     SOCS6, TBXAS1, AC012150.1, AC008691.1, OLR1, P2RY12, IRAK3, LINC02798, SLC4A7, AC079015.1
## PC_ 3
## Positive:  NXPH1, GRIP1, KCNIP1, GRIK1, SYTL5, GAD1, ERBB4, GAD2, VWC2, PTPRM
##     KIF26B, LINC01322, KIAA1211, NHS, RBMS3, GRIK2, SOX6, UBASH3B, SLC24A3, IGF1
##     SNRPN, ROBO1, PTCHD4, GRIP2, SGCZ, TOX2, ANK1, NTN4, XKR4, DCX
## Negative:  SATB2, LY86-AS1, USP43, ENC1, SLIT3, AC011287.1, NELL2, MCHR2, ANO3, ADAMTS16
##     MLIP, KIAA1211L, R3HDM1, EPHA4, LMO4, PTPRD, SV2B, NRGN, CCBE1, LINC02217
##     LINC01250, AC011995.2, PTK2B, STYK1, CELF2, NGEF, LDB2, SMAD3, CBLN2, LRRK1
## PC_ 4
## Positive:  SMOC1, LINC01505, BCAS1, XYLT1, CA10, TNR, AC008571.2, SEMA3C, NECAB1, AL589693.1
##     MEGF11, SOX2-OT, PDGFRA, LRRC4C, RASGRF2, PCDH15, LHFPL3, AL512308.1, SGCD, NECTIN3
##     AL391840.1, MYT1, NLGN4X, GRIA2, LRRN1, GRIA4, AC004852.2, COL9A1, PPFIA2, LRRC7
## Negative:  LRMDA, FKBP5, ARHGAP24, SRGAP2B, PLXDC2, DOCK8, IRAK3, RGS1, ST6GAL1, SMAP2
##     PRKCH, LINC01374, PDE3B, LYN, IPCEF1, LPCAT2, PALD1, APBB1IP, PTPRC, ABCC4
##     TBXAS1, RUNX1, ATP8B4, ADAM28, PIK3R5, TGFBR2, ARHGAP15, LNCAROD, OLR1, IKZF1
## PC_ 5
## Positive:  SYTL5, KIRREL3, SNRPN, NRXN3, TMTC1, FAM135B, MARCH11, GRIP1, TENM2, MARCH4
```

```
##       MYO5B, PTCHD4, TENM3, GAD2, RASGRF2, ADAMTS19, AL121578.3, SNHG14, HCN1, ELAVL2
##       CDH9, ROBO1, KCNC2, CDH8, GRIA1, DAPK1, PACRG, GRIP2, LINC02389, IGF1
## Negative:  TNR, MEGF11, PTPRZ1, XYLT1, LHFPL3, PCDH15, PDGFRA, DSCAM, OPHN1, VCAN
##       AC004852.2, SLC35F1, BX284613.2, MYT1, CA10, LUZP2, STK32A, COL11A1, COL9A1, AFAP1L2
##       SEMA5A, CRISPLD2, MMP16, SMOC1, AL512308.1, FERMT1, PLAAT1, ETV1, LRRK2, AC124254.2

## Normalizing layer: counts.Gene Expression

## Finding variable features for layer counts.Gene Expression

## Centering and scaling data matrix

## PC_ 1
## Positive:  SCN2A, KALRN, CACNA1C, GRIN2B, SCN8A, KSR2, DLGAP2, MYT1L, SLC4A10, GABBR2
##       GABRB3, AGBL4, STXBP5L, CELF2, MYH10, TMEM132B, PTPRN2, CERS6, DYNC1I1, VPS13A
##       UNC80, MAP2, RIMS2, SNAP25, KCNMA1, ERC2, AEBP2, RGS7, LRFN5, CACNA1B
## Negative:  BCAS1, AC008571.2, COLEC12, SLC14A1, AP000477.3, AC012405.1, RGR, AL627316.1, AC026316.5,
##       ETNPPL, SLCO1C1, CD44, AL137139.2, AC073941.1, HNF4G, FAM189A2, SLC39A12, SERPINA3, AC114485.1
##       TNC, DIPK2B, GLI3, F3, MRVI1, AL162725.2, BBOX1, SPP1, AQP4, AL591115.1
## PC_ 2
## Positive:  NXPH1, GRIP1, GAD1, KCNIP1, PTPRM, PTCHD4, ERBB4, NHS, UBASH3B, VWC2
##       GAD2, KIAA1211, SOX6, ANK1, SYTL5, SNRPN, DOCK11, GRIK1, PAM, WLS
##       KCNC2, IGF1, ZNF385D, LINC01322, GRIP2, BTBD11, ADCY8, KLF12, GRIK3, NRIP3
## Negative:  AC008574.1, TAFA1, AC011287.1, CUX2, LINC02055, STYK1, LY86-AS1, CNGB1, AP000829.1, ENC1
##       AL157944.1, MCHR2, AL583827.1, LINC01378, ADAMTS3, USP43, TTC6, AL024495.1, KCNG3, R3HDM1
##       TESPA1, HS6ST3, LINC01331, FAT3, RGS6, CYP1B1-AS1, MGAT5B, LRRC2, KIAA2012, LDB2
## PC_ 3
## Positive:  FKBP5, ST6GAL1, LRMDA, MERTK, IRAK3, PRKCH, NHSL1, TGFBR2, PALD1, LYN
##       RGS1, IPCEF1, ARHGAP24, APBB1IP, DOCK8, IL13RA1, LINC01374, TBXAS1, SP100, LPCAT2
##       PIK3R5, LNCAROD, PTPRC, SLCO2B1, FLI1, SRGN, SOCS6, OLR1, ADAM28, RUNX1
## Negative:  NXPH1, GAD1, KCNC2, GAD2, SYTL5, ERBB4, ROBO1, RAB3C, GRIP1, ANK1
##       GRIK1, VWC2, KIRREL3, STXBP6, LINC01322, SCN1A-AS1, BTBD11, MYO16, AL121578.3, PTCHD4
##       KCNMB2, GPC6, KIF26B, NHS, GRIK2, SGCZ, AC087457.1, GRIP2, SNRPN, TENM3
## PC_ 4
## Positive:  HS3ST4, FOXP2, PDZRN4, RXFP1, HS3ST2, TLE4, SORCS3, PCDH11Y, LINC02232, CLSTN2
##       AL353709.1, FRMPD4, AP000943.2, PTK7, DIRAS2, PCDH11X, KIAA1217, MSC-AS1, SLC8A1, SLC35F1
##       ADTRP, VWA2, ARPP21, LINC02223, LINC02082, SLC35F3, SEMA3E, B3GALT2, TRMT9B, SYT6
## Negative:  GLIS3, TPD52L1, RFX4, ACSS3, AL137139.2, SLC14A1, BMPR1B, AC012405.1, AQP4, SHROOM3
##       WDR49, COLEC12, MRVI1, CD44, AC073941.1, FAM189A2, FAM107A, ETNPPL, SLC7A11, AC002429.2
##       MIR4300HG, IGFBP7, GJA1, SLC4A4, ANGPT1, SLCO1C1, PAMR1, GLI3, AL606923.2, RANBP3L
## PC_ 5
## Positive:  RGS1, LINC01374, LPCAT2, DOCK8, APBB1IP, LYN, PIK3R5, PALD1, IRAK3, OLR1
##       PTPRC, LNCAROD, PRKCH, TBXAS1, ADAM28, PLA2G4A, SOCS6, CD53, PDE3B, P2RY12
##       AC012150.1, IKZF1, SYK, TGFBR2, IL13RA1, AC079015.1, TMEM156, CSF2RA, IPCEF1, CPED1
## Negative:  RFX4, AL137139.2, SLC14A1, AC012405.1, TPD52L1, ACSS3, LINC00299, AQP4, BMPR1B, WDR49
##       CD44, AC073941.1, MRVI1, PCSK5, FAM189A2, AC002429.2, ETNPPL, COLEC12, AC110023.1, COL21A1
##       CDHR3, BBOX1, SLCO1C1, HS3ST4, GJA1, TRMT9B, ANGPT1, TNC, AQP4-AS1, PDZRN4
```

```r
print("PCA done")
```

```
## [1] "PCA done"
```

```r
# 2) Merge all samples into one Seurat object (RNA + ATAC assays carried along)
merged <- merge(
  x           = sample_seurat_list[[1]],
  y           = sample_seurat_list[-1],
  add.cell.ids = names(sample_seurat_list) ,
```

```
  project      = "MergedProject"
)

print("Merged done")
```

[1] "Merged done"

```
# 3) Re-run PCA on the merged RNA (just in case)
DefaultAssay(merged) <- "RNA"
merged <- merged %>%
  NormalizeData(assay = "RNA", layer = "counts.Gene Expression") %>%
  FindVariableFeatures(assay = "RNA", layer = "counts.Gene Expression",
                       nfeatures = 3000) %>%
  ScaleData(assay = "RNA", layer = "counts.Gene Expression") %>%
  RunPCA(assay = "RNA", reduction.name = "pca", npcs = 50)
```

Normalizing layer: counts.Gene Expression.CTRL

Normalizing layer: counts.Gene Expression.TSC_edge

Normalizing layer: counts.Gene Expression.TSC_tuber

Normalizing layer: counts.Gene Expression.TSC_outside

Finding variable features for layer counts.Gene Expression.CTRL

Finding variable features for layer counts.Gene Expression.TSC_edge

Finding variable features for layer counts.Gene Expression.TSC_tuber

Finding variable features for layer counts.Gene Expression.TSC_outside

Centering and scaling data matrix

PC_ 1
Positive: FAM155A, RBFOX1, KALRN, LRRC7, ANKS1B, SCN8A, SCN2A, GABRB3, MYT1L, UNC79
GRIN2B, KSR2, ARHGAP32, CACNA1C, AGBL4, STXBP5L, SLC4A10, MYH10, KCNMA1, CELF2
SNHG14, MSRA, CSMD1, GABBR2, UNC80, SNAP25, ADGRB3, PTPRN2, ATP8A2, SNAP91
Negative: LINC01170, AC008571.2, BCAS1, SPP1, LINC01445, ENPP6, DOCK8, AP000477.3, PTPRC, RGS1
PDK4, PIK3R5, FYB1, IKZF1, AC114485.1, IL13RA1, LNCAROD, APBB1IP, TBXAS1, VSIG4
SYK, TLR2, TGFBR2, TMEM156, ITGAX, CSF2RA, TFEC, AL591115.1, AL392172.2, SP100
PC_ 2
Positive: SLC8A2, KCNH4, AL117329.1, RTN4RL1, CAMK2A, COL26A1, CUX2, IGSF21, PDE2A, AC011995.2
CABP1, MEIS3, CELF5, CHD5, DGCR9, SYT7, CAMKK1, HRH2, NWD2, CDH22
RBFOX3, KCNQ2, GRIN1, DLGAP2, LINC01170, SLC22A10, ENC1, NRGN, SLC17A7, SMAP2
Negative: TPD52L1, RFX4, AL137139.2, ACSS3, STON2, SLC14A1, AC012405.1, GPM6A, AQP4, GLIS3
BMPR1B, PLCE1, PLSCR4, MRVI1, RYR3, WDR49, CD44, PRKG1, ARHGEF26, AC073941.1
LINC01748, COLEC12, RNF19A, AC091826.2, SLC16A9, FAM189A2, SHROOM3, AC099792.1, AC114971.1, FAM10
PC_ 3
Positive: TAFA1, CUX2, MCHR2, AC011287.1, ENC1, RFTN1, TESPA1, AP000829.1, AL583827.1, R3HDM1
LY86-AS1, MED12L, CDH22, LINC00507, AC008574.1, DLGAP1-AS4, LINC01331, SMAD3, STYK1, NWD2
KCNH4, TTC6, SNED1, LINC01250, AC011995.2, MAPK4, USP43, NGEF, RGS6, GLIS3
Negative: GRIP1, NXPH1, GAD1, SNRPN, KCNIP1, UBASH3B, NHS, PTPRM, GAD2, SYTL5
KIAA1211, VWC2, ERBB4, IGF1, ANK1, GRIK1, PTCHD4, DOCK11, GRIP2, ZNF385D
KCNC2, GRIN2D, NRIP3, NTN4, GRIK2, BTBD11, KIF26B, ROBO1, XKR4, AC087457.1
PC_ 4
Positive: SOX2-OT, ERBB4, GAD1, NXPH1, LINC01170, LHFPL3, GRIP1, GALNT13, GRIK1, PCDH15
SLC24A3, TMEFF2, KIAA1211, ROBO1, VWC2, SNRPN, GAD2, SOX6, SYTL5, BCHE
GRIN2D, GRIP2, LINC01322, ANK1, KIF26B, KAZN, SMOC1, GRIK2, PLD5, NRIP3
```

```
## Negative:  LRMDA, DOCK8, FKBP5, IRAK3, APBB1IP, ARHGAP24, SMAP2, ST6GAL1, PALD1, LINC01374
##     LYN, TBXAS1, LPCAT2, PRKCH, RGS1, IPCEF1, ADAM28, IL13RA1, TGFBR2, LNCAROD
##     PIK3R5, PTPRC, ARHGAP15, OLR1, P2RY12, AC012150.1, ATP8B4, CSF2RA, SYK, PDE3B
## PC_ 5
## Positive:  PDZRN4, RXFP1, FOXP2, HS3ST4, HS3ST2, CLSTN2, PTK7, AL353709.1, LINC02232, TRMT9B
##     DPY19L1, DPP10, TLE4, SLC35F1, FAM189A1, MDGA2, SORCS3, GABRA5, POU6F2, NPTX1
##     GRIK3, AC008415.1, NELL2, DIRAS2, FRMPD4, AP000943.2, AL138701.2, SSTR2, KHDRBS3, SLIT1
## Negative:  EPHA6, CUX2, GRIA4, PDZD2, LINC01331, NDST3, PIP5K1B, LAMB1, DAB1, SLIT2
##     NECTIN3, GPC6, AC008574.1, CYP1B1-AS1, TTC6, ADAMTS9-AS2, HS6ST3, KCNMB2, VSTM2A, AL583827.1
##     VSTM2A-OT1, RASGRF2, TAFA1, THSD7A, PRKG2, LRRTM4, CPNE8, LINC00507, KIRREL3, EPB41L3
```

```r
print("PCA for all is done")
```

```
## [1] "PCA for all is done"
```

```r
# 4) Harmony integration on the RNA PCs, correcting for your batch/group
merged <- RunHarmony(
  object        = merged,
  group.by.vars = "orig.ident",
  reduction.use = "pca",
  dims.use      = 1:50
)
```

```
## Transposing data matrix
```

```
## Initializing state using k-means centroids initialization
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 1969400)
```

```
## Warning: did not converge in 25 iterations
## Warning: did not converge in 25 iterations
```

```
## Harmony 1/10
```

```
## Harmony 2/10
```

```
## Harmony 3/10
```

```
## Harmony 4/10
```

```
## Harmony 5/10
```

```
## Harmony 6/10
```

```
## Harmony 7/10
```

```
## Harmony converged after 7 iterations
```

```r
print("Harmony done")
```

```
## [1] "Harmony done"
```

```r
# 5) UMAP & clustering on the harmony embedding
merged <- merged %>%
  RunUMAP(reduction = "harmony", dims = 1:50) %>%
  FindNeighbors(reduction = "harmony", dims = 1:50) %>%
  FindClusters(resolution = 0.5)
```

```
## Warning: The default method for RunUMAP has changed from calling Python UMAP via reticulate to the R-
## To use Python UMAP via reticulate, set umap.method to 'umap-learn' and metric to 'correlation'
## This message will be shown once per session
```

```
## 21:23:26 UMAP embedding parameters a = 0.9922 b = 1.112
```

```
## 21:23:26 Read 39388 rows and found 50 numeric columns

## 21:23:26 Using Annoy for neighbor search, n_neighbors = 30

## 21:23:26 Building Annoy index with metric = cosine, n_trees = 50

## 0%   10   20   30   40   50   60   70   80   90   100%

## [----|----|----|----|----|----|----|----|----|----|

## **************************************************|
## 21:23:29 Writing NN index file to temp file /var/folders/62/9y36k0q92vv6hckwc077f_1r0000gn/T//Rtmp8E
## 21:23:29 Searching Annoy index using 1 thread, search_k = 3000
## 21:23:37 Annoy recall = 98.42%
## 21:23:38 Commencing smooth kNN distance calibration using 1 thread with target n_neighbors = 30
## 21:23:38 Initializing from normalized Laplacian + noise (using RSpectra)
## 21:23:41 Commencing optimization for 200 epochs, with 1843866 positive edges
## 21:23:41 Using rng type: pcg
## 21:23:52 Optimization finished
## Computing nearest neighbor graph

## Warning: package 'future' was built under R version 4.4.1

## Computing SNN

## Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan van Eck
##
## Number of nodes: 39388
## Number of edges: 1702737
##
## Running Louvain algorithm...
## Maximum modularity in 10 random starts: 0.9336
## Number of communities: 24
## Elapsed time: 6 seconds
```

```r
print("UMAP and clustering done")
```

```
## [1] "UMAP and clustering done"
```

```r
# visualize RNA integration
p<-DimPlot(merged, reduction = "umap", group.by = "orig.ident") +
  ggtitle("RNA Harmony UMAP")
ggsave(p, filename = "./results/Merge-RNA-analysis-plot/RNA_Harmony_UMAP.png",
       width = 8, height = 6, dpi = 300)

p<-DimPlot(merged, reduction = "umap", group.by = "orig.ident",
          split.by = "orig.ident") + ggtitle("RNA Harmony UMAP")
ggsave(p, filename = "./results/Merge-RNA-analysis-plot/RNA_Harmony_UMAP_split.png",
       width = 20, height = 5, dpi = 300)

# show the cluster results
p<-DimPlot(merged, reduction = "umap", group.by = "seurat_clusters") +
  ggtitle("RNA Harmony UMAP Clusters")
ggsave(p, filename = "./results/Merge-RNA-analysis-plot/RNA_Harmony_UMAP_clusters.png",
       width = 8, height = 6, dpi = 300)


DefaultAssay(merged) <- "RNA"
```

```r
# save the data
saveRDS(merged, here::here("data", "processed-data","merged_RNA_2025-05-06.rds"),
        compress = FALSE)
```