GSVA data preparation

```
library(Seurat)
library(GSEABase)
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

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

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

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
## Loading required package: annotate
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: stats4
```

```
## Loading required package: IRanges

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:base':
##
##     expand.grid

## Loading required package: XML

## Loading required package: graph

##
## Attaching package: 'graph'

## The following object is masked from 'package:XML':
##
##     addNode
```

```
dir.create('../GSVA_flt_res0.3')
setwd('../GSVA_flt_res0.3')
dir.create('tmp')
dir.create('gsva')

subdataset <- readRDS('../tmp/crc_smc.malignantcells.Rds')
subdataset <- subset(subdataset, cells = rownames(subset(subdataset@meta.data, !RNA_snn_res.0.3 %in% c(
subdataset@meta.data <- droplevels(subdataset@meta.data)

subdataset@meta.data$RNA_snn_res.0.3 <- factor(subdataset@meta.data$RNA_snn_res.0.3, levels = c(0:6))
head(subdataset@meta.data); nrow(subdataset@meta.data) # 17276
```

```
##                                   nCount_RNA nFeature_RNA         Library
## AAACCTGCATACGCCG-1-PM-PS-0001-T-A1      35998         4823 PM-PS-0001-T-A1
## AAACCTGGTCGCATAT-1-PM-PS-0001-T-A1      31383         5252 PM-PS-0001-T-A1
## AAACCTGTCCCTTGCA-1-PM-PS-0001-T-A1       7302         1713 PM-PS-0001-T-A1
## AAACGGGAGGGAAACA-1-PM-PS-0001-T-A1       3759         1233 PM-PS-0001-T-A1
## AAACGGGGTATAGGTA-1-PM-PS-0001-T-A1      23097         3874 PM-PS-0001-T-A1
## AAAGATGAGGCCGAAT-1-PM-PS-0001-T-A1      14860         3282 PM-PS-0001-T-A1
##                                   Patient  Sample Cell_subtype RNA_snn_res.0.3
## AAACCTGCATACGCCG-1-PM-PS-0001-T-A1   SMC01 SMC01-T         CMS2               2
## AAACCTGGTCGCATAT-1-PM-PS-0001-T-A1   SMC01 SMC01-T         CMS2               0
## AAACCTGTCCCTTGCA-1-PM-PS-0001-T-A1   SMC01 SMC01-T         CMS2               2
## AAACGGGAGGGAAACA-1-PM-PS-0001-T-A1   SMC01 SMC01-T         CMS2               1
## AAACGGGGTATAGGTA-1-PM-PS-0001-T-A1   SMC01 SMC01-T         CMS2               1
## AAAGATGAGGCCGAAT-1-PM-PS-0001-T-A1   SMC01 SMC01-T         CMS2               2
```

```
## [1] 17276
```

```
summary(subdataset@meta.data$RNA_snn_res.0.3)
```

```
##    0    1    2    3    4    5    6
## 6609 4550 3378 1372  807  325  235
```

```
label <- subdataset@meta.data
data <- data.frame(as.matrix(GetAssayData(subdataset, slot = 'data', assay = 'RNA')), check.rows = F, ch
data[1:4, 1:4]
```

```
##                              AAACCTGCATACGCCG-1-PM-PS-0001-T-A1
## TSPAN6-ENSG00000000003.15-5                          0.4418526
## TNMD-ENSG00000000005.6-4                             0.0000000
## DPM1-ENSG00000000419.12-6                            1.0799568
## SCYL3-ENSG00000000457.14-7                           0.0000000
##                              AAACCTGGTCGCATAT-1-PM-PS-0001-T-A1
## TSPAN6-ENSG00000000003.15-5                          0.6708666
## TNMD-ENSG00000000005.6-4                             0.0000000
## DPM1-ENSG00000000419.12-6                            1.1726391
## SCYL3-ENSG00000000457.14-7                           0.2766038
##                              AAACCTGTCCCTTGCA-1-PM-PS-0001-T-A1
## TSPAN6-ENSG00000000003.15-5                          0.0000000
## TNMD-ENSG00000000005.6-4                             0.0000000
## DPM1-ENSG00000000419.12-6                            0.8626738
## SCYL3-ENSG00000000457.14-7                           0.0000000
##                              AAACGGGAGGGAAACA-1-PM-PS-0001-T-A1
## TSPAN6-ENSG00000000003.15-5                                  0
## TNMD-ENSG00000000005.6-4                                     0
## DPM1-ENSG00000000419.12-6                                    0
## SCYL3-ENSG00000000457.14-7                                   0
```

```
### Pseudo-bulk transformation ###
pseudo_data <- data.frame(row.names = rownames(data), matrix(nrow = length(rownames(data)), ncol = leng
colnames(pseudo_data) <- levels(label$RNA_snn_res.0.3)
head(pseudo_data)
```

```
##                               0  1  2  3  4  5  6
## TSPAN6-ENSG00000000003.15-5  NA NA NA NA NA NA NA
## TNMD-ENSG00000000005.6-4     NA NA NA NA NA NA NA
## DPM1-ENSG00000000419.12-6    NA NA NA NA NA NA NA
## SCYL3-ENSG00000000457.14-7   NA NA NA NA NA NA NA
## C1orf112-ENSG00000000460.17-7 NA NA NA NA NA NA NA
## FGR-ENSG00000000938.13-6     NA NA NA NA NA NA NA
```

```
for (clst in levels(label$RNA_snn_res.0.3)) {
  use_bc <- rownames(subset(label, RNA_snn_res.0.3 == clst))
  pseudo_data[, clst] <- rowMeans(data[, use_bc])
}
head(pseudo_data)
```

```
##                                          0           1           2           3
## TSPAN6-ENSG00000000003.15-5   0.503763546 0.430795073 0.496466656 0.2691926063
```

```
## TNMD-ENSG00000000005.6-4      0.008381026 0.019345066 0.009537888 0.0058047225
## DPM1-ENSG00000000419.12-6     0.576895933 0.505894665 0.705633088 0.4036800172
## SCYL3-ENSG00000000457.14-7    0.027537204 0.028486198 0.028321173 0.0091273321
## C1orf112-ENSG00000000460.17-7 0.011633786 0.008446574 0.053402234 0.0144090395
## FGR-ENSG00000000938.13-6      0.001862344 0.001989247 0.001241530 0.0008870705
##                                          4           5           6
## TSPAN6-ENSG00000000003.15-5   0.3874863503 0.03779613 0.508128726
## TNMD-ENSG00000000005.6-4      0.0005113429 0.00000000 0.003130667
## DPM1-ENSG00000000419.12-6     0.4604125478 0.15446692 0.470009515
## SCYL3-ENSG00000000457.14-7    0.0327217456 0.03071259 0.024169450
## C1orf112-ENSG00000000460.17-7 0.0098759336 0.02585296 0.017580665
## FGR-ENSG00000000938.13-6      0.0032123432 0.00000000 0.001237696
```

```
remove(data)

symbols <- unlist(lapply(rownames(pseudo_data), function(x) unlist(strsplit(as.character(x), split = '-
pseudo_data <- data.frame(Symbol = symbols, pseudo_data, check.rows = F, check.names = F)
pseudo_data[1:4, 1:4]
```

```
##                               Symbol           0          1           2
## TSPAN6-ENSG00000000003.15-5 TSPAN6 0.503763546 0.43079507 0.496466656
## TNMD-ENSG00000000005.6-4      TNMD 0.008381026 0.01934507 0.009537888
## DPM1-ENSG00000000419.12-6     DPM1 0.576895933 0.50589467 0.705633088
## SCYL3-ENSG00000000457.14-7   SCYL3 0.027537204 0.02848620 0.028321173
```

```
### Redundant gene selection ###
duplicated <- data.frame(table(symbols)); remove(symbols)
duplicated <- duplicated[duplicated$Freq > 1, ]
duplicated
```

```
##           symbols Freq
## 81          ABCF2    2
## 639    AC005618.1    2
## 1319   AC008731.1    2
## 1449   AC009065.2    2
## 1593   AC009495.1    2
## 1977   AC011453.1    2
## 2025   AC011498.1    2
## 2410   AC016586.1    2
## 2752   AC021078.1    2
## 2921   AC022558.1    2
## 4232   AC087269.1    2
## 4364   AC090186.1    2
## 4418   AC090559.2    2
## 4870   AC093157.1    2
## 4984   AC093788.1    2
## 5110   AC097382.1    2
## 5124   AC097493.1    2
## 6513   AC135983.2    2
## 7268   AF131215.5    2
## 7387         AHRR    2
## 7685   AL031602.1    2
## 7982   AL109615.1    2
```

```
## 8036   AL117209.1    2
## 8341   AL136115.1    2
## 8410   AL136987.1    2
## 8430   AL137127.1    2
## 8985   AL353898.1    2
## 9780   AL513327.1    2
## 9869   AL590226.1    2
## 9925   AL590762.1    2
## 10084  AL627309.1    2
## 11108  AP006222.2    2
## 11287   ARHGAP11B    2
## 12547    C2orf27A    2
## 17524        GGT1    2
## 18899      HSPA14    2
## 21015   LINC01238    2
## 22469       MATR3    2
## 22659 Metazoa-SRP    7
## 26103     POLR2J3    2
## 26104     POLR2J4    2
## 27650    RN7SKP23    2
## 29936    SCARNA11    2
## 31155     SNORA11    3
## 31163      SNORA2    2
## 31165     SNORA22    2
## 31167     SNORA24    2
## 31168     SNORA25    4
## 31169     SNORA26    3
## 31173     SNORA31    4
## 31180     SNORA42    3
## 31182     SNORA48    2
## 31187     SNORA57    2
## 31195     SNORA67    2
## 31198      SNORA7    2
## 31201     SNORA72    2
## 31202     SNORA73    3
## 31204     SNORA74    2
## 31207     SNORA75    2
## 31210     SNORA79    2
## 31214     SNORA81    2
## 31242     snoU109    2
## 31243      snoU13   40
## 31360        SOD2    2
## 32261        TBCE    2
## 33788          U3   11
## 33796          U6    2
## 34624       Y-RNA  104
```

```
selected_iso <- c()

for (dup in duplicated$symbols) {
  tmp <- pseudo_data[pseudo_data$Symbol == dup, ]
  selected <- names(sort(rowSums(tmp[, -1]), decreasing = T)[1])
  selected_iso <- append(selected_iso, selected)
  remove(tmp)
```

```
}

iso_pseudo_data <- pseudo_data[selected_iso, ]
iso_pseudo_data[1:4, 1:4]; nrow(iso_pseudo_data)
```

```
##                                    Symbol            0            1
## ABCF2-ENSG00000033050.9-6           ABCF2 1.411129e-01 0.1044621659
## AC005618.1-ENSG00000272070.1-6 AC005618.1 6.845838e-04 0.0009707298
## AC008731.1-ENSG00000256439.1   AC008731.1 1.030248e-03 0.0012441804
## AC009065.2-ENSG00000207715.1   AC009065.2 4.786005e-05 0.0000000000
##                                         2
## ABCF2-ENSG00000033050.9-6      0.1799957561
## AC005618.1-ENSG00000272070.1-6 0.0003465819
## AC008731.1-ENSG00000256439.1   0.0006629246
## AC009065.2-ENSG00000207715.1   0.0002496389
```

```
## [1] 68
```

```
pseudo_data <- subset(pseudo_data, Symbol %in% setdiff(pseudo_data$Symbol, duplicated$symbols) )
pseudo_data <- rbind(pseudo_data, iso_pseudo_data); remove(iso_pseudo_data)
head(pseudo_data); nrow(pseudo_data)
```

```
##                                  Symbol           0           1           2
## TSPAN6-ENSG00000000003.15-5     TSPAN6 0.503763546 0.430795073 0.496466656
## TNMD-ENSG00000000005.6-4          TNMD 0.008381026 0.019345066 0.009537888
## DPM1-ENSG00000000419.12-6         DPM1 0.576895933 0.505894665 0.705633088
## SCYL3-ENSG00000000457.14-7       SCYL3 0.027537204 0.028486198 0.028321173
## C1orf112-ENSG00000000460.17-7 C1orf112 0.011633786 0.008446574 0.053402234
## FGR-ENSG00000000938.13-6           FGR 0.001862344 0.001989247 0.001241530
##                                       3           4          5           6
## TSPAN6-ENSG00000000003.15-5   0.2691926063 0.3874863503 0.03779613 0.508128726
## TNMD-ENSG00000000005.6-4      0.0058047225 0.0005113429 0.00000000 0.003130667
## DPM1-ENSG00000000419.12-6     0.4036800172 0.4604125478 0.15446692 0.470009515
## SCYL3-ENSG00000000457.14-7    0.0091273321 0.0327217456 0.03071259 0.024169450
## C1orf112-ENSG00000000460.17-7 0.0144090395 0.0098759336 0.02585296 0.017580665
## FGR-ENSG00000000938.13-6      0.0008870705 0.0032123432 0.00000000 0.001237696
```

```
## [1] 35690
```

```
remove(subdataset)

rownames(pseudo_data) <- pseudo_data$Symbol
pseudo_data$Symbol <- NULL

### Filter low expression genes ###
pseudo_data$Sums <- rowSums(pseudo_data)
use_gene <- rownames(subset(pseudo_data, Sums > 0.01)); length(use_gene)
```

```
## [1] 18723
```

```
pseudo_data$Sums <- NULL

for (sample in colnames(pseudo_data)) {
  add_gene <- setdiff(rownames(pseudo_data[pseudo_data[, sample] > 0.01, ]), use_gene)
  use_gene <- c(use_gene, add_gene)
}
pseudo_data <- pseudo_data[use_gene, ]
colnames(pseudo_data) <- unlist(lapply(colnames(pseudo_data), function(x) paste0(c('res0.3_', as.charac
head(pseudo_data); nrow(pseudo_data) # 18723
```

```
##             res0.3_0    res0.3_1    res0.3_2     res0.3_3     res0.3_4
## TSPAN6   0.503763546 0.430795073 0.496466656 0.2691926063 0.3874863503
## TNMD     0.008381026 0.019345066 0.009537888 0.0058047225 0.0005113429
## DPM1     0.576895933 0.505894665 0.705633088 0.4036800172 0.4604125478
## SCYL3    0.027537204 0.028486198 0.028321173 0.0091273321 0.0327217456
## C1orf112 0.011633786 0.008446574 0.053402234 0.0144090395 0.0098759336
## FGR      0.001862344 0.001989247 0.001241530 0.0008870705 0.0032123432
##             res0.3_5    res0.3_6
## TSPAN6   0.03779613 0.508128726
## TNMD     0.00000000 0.003130667
## DPM1     0.15446692 0.470009515
## SCYL3    0.03071259 0.024169450
## C1orf112 0.02585296 0.017580665
## FGR      0.00000000 0.001237696
```

```
## [1] 18723
```

```
pData <- AnnotatedDataFrame(data.frame(row.names = colnames(pseudo_data), clustering = c(0:6)))
exprSet <- ExpressionSet(as.matrix(pseudo_data), phenoData = pData, annotation = 'Symbol'); exprSet
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 18723 features, 7 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: res0.3_0 res0.3_1 ... res0.3_6 (7 total)
##   varLabels: clustering
##   varMetadata: labelDescription
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation: Symbol
```

```
saveRDS(exprSet, 'tmp/exprSet.Rds')
```