Figure 4

EmptyNN - Figure 4

The following code reproduces the Figure 4 in our EmptyNN manuscript.

Load libraries

```r
# Please download datasets and seurat objects before running this analysis (run download_datasets.sh in
library("Matrix")
```

```
## Warning: package 'Matrix' was built under R version 4.0.5
```

```r
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```r
# R version 3.6.3, Matrix_1.3-2, ggplot_2_3.3.3
```

Load data

```r
raw_counts <- readRDS("../data/mouse_nuclei_2k_raw.rds")
retained <- readRDS("../data/mouse_nuclei_retained.rds")
```

Figure 4A

```r
# Define ambient RNA sigature ####
n_counts <- colSums(raw_counts)
p_set <- names(n_counts[n_counts<100 & n_counts>10])
p_set <- sample(p_set,2000)
p_set <- raw_counts[,p_set]
freq <- apply(p_set,1,function(x) sum(x>0))
ambient <- names(freq)[tail(order(freq),100)]
# EmptyNN
df <- retained@reductions$tsne@cell.embeddings
amb <- ambient[ambient %in% rownames(retained)]
ambient_expr <- colSums(retained@assays$RNA@counts[amb,])
df <- data.frame(df,"ambient_expr"=ambient_expr,retained@meta.data)
```

```
df$method <- "Both"
df$method[df$emptynn & !df$cellrangerv3] <- "EmptyNN"
df$method[!df$emptynn & df$cellrangerv3] <- "CellRanger_v3"
df1 <- df
df1$ambient_expr[df1$method!='EmptyNN'] <- 0
df1$method2 <- "EmptyNN"
df2 <- df
df2$ambient_expr[df2$method!='CellRanger_v3'] <- 0
df2$method2 <- "CellRanger_v3"
tmp <- rbind(df1,df2)
ggplot(tmp,aes(tSNE_1,tSNE_2,color=log(ambient_expr+1)))+facet_wrap(~method2)+
    geom_point(aes(alpha=log(ambient_expr+1)))+theme_bw()+
    scale_color_gradient2(low="grey",mid="lightblue",high="darkblue",midpoint = 3)+
    ggtitle("Ambient RNA Expression")
```
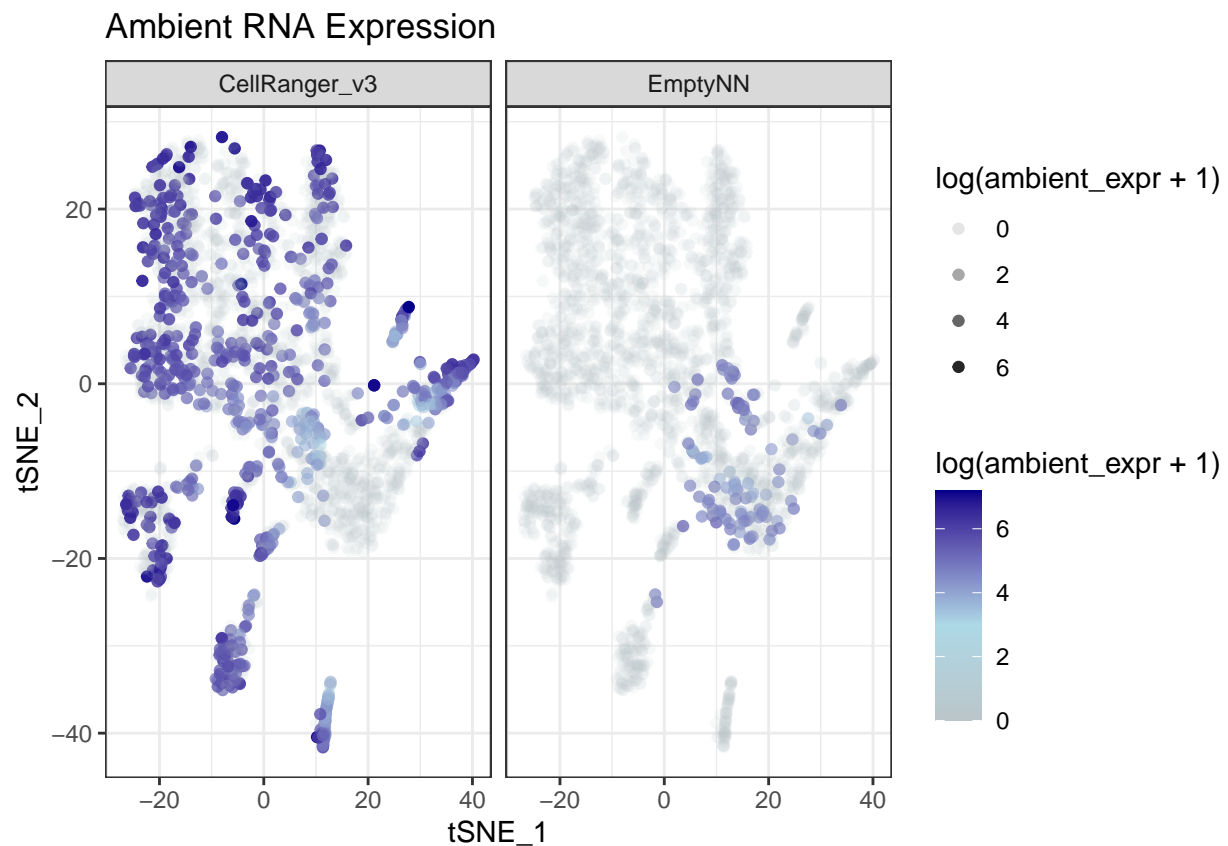

Figure 4B

```
# sliced ratio ####
spliced <- t(read.delim("../data/spliced.csv",sep="\t",header = T,row.names = 1))
unspliced <- t(read.delim("../data/unspliced.csv",sep="\t",header = T,row.names = 1))
s <- colSums(spliced)
u <- colSums(unspliced)
```

```
p <- data.frame(s,u)
rownames(p) <- gsub("possorted_genome_bam_HWVYF:","",colnames(spliced))
rownames(p) <- gsub("x","-1",rownames(p))
p$pcentage <- apply(p,1,function(x){x[1]/sum(x)})

retained$spliced_ratio <- p[match(colnames(retained),rownames(p)),'pcentage']

df <- retained@reductions$tsne@cell.embeddings
df <- data.frame(df,retained@meta.data)
df$method <- "Both"
df$method[df$emptynn & !df$cellrangerv3] <- "EmptyNN"
df$method[!df$emptynn & df$cellrangerv3] <- "cellRangerv3"
df1 <- df
df1$spliced_ratio[df1$method!='EmptyNN'] <- 0
df1$method2 <- "EmptyNN"
df2 <- df
df2$spliced_ratio[df2$method!='cellRangerv3'] <- 0
df2$method2 <- "cellRangerv3"
tmp <- rbind(df1,df2)
ggplot(tmp,aes(tSNE_1,tSNE_2,color=spliced_ratio))+facet_wrap(~method2)+
  geom_point(aes(alpha=spliced_ratio))+theme_bw()+
  scale_color_gradient2(low="grey",mid="lightblue",high="darkblue",midpoint = 0.25)+
  ggtitle("Percentage spliced reads")
```

Figure 4C

```
a <- retained[,retained$emptynn]$spliced_ratio
df1 <- data.frame("expr"=a,"method"='EmptyNN')
b <- retained[,retained$cellrangerv3]$spliced_ratio
df2 <- data.frame("expr"=b,"method"='CellRanger_v3')
c <- retained[,retained$cellbender]$spliced_ratio
df3 <- data.frame("expr"=c,"method"='CellBender')
d <- retained[,retained$emptydrops]$spliced_ratio
df4 <- data.frame("expr"=d,"method"='EmptyDrops')
e <- retained[,retained$cellrangerv2]$spliced_ratio
df5 <- data.frame("expr"=b,"method"='CellRanger_v2')
tmp <- rbind(df1,df2,df3,df4,df5)
#ecdf plot
ggplot(tmp, aes(x=expr, colour=method,linetype=method),size=5) +
  xlab("Spliced/Total ratio")+theme_bw()+
  stat_ecdf()+ylab("Percentile")
```

```
## Warning: Removed 187 rows containing non-finite values (stat_ecdf).
```
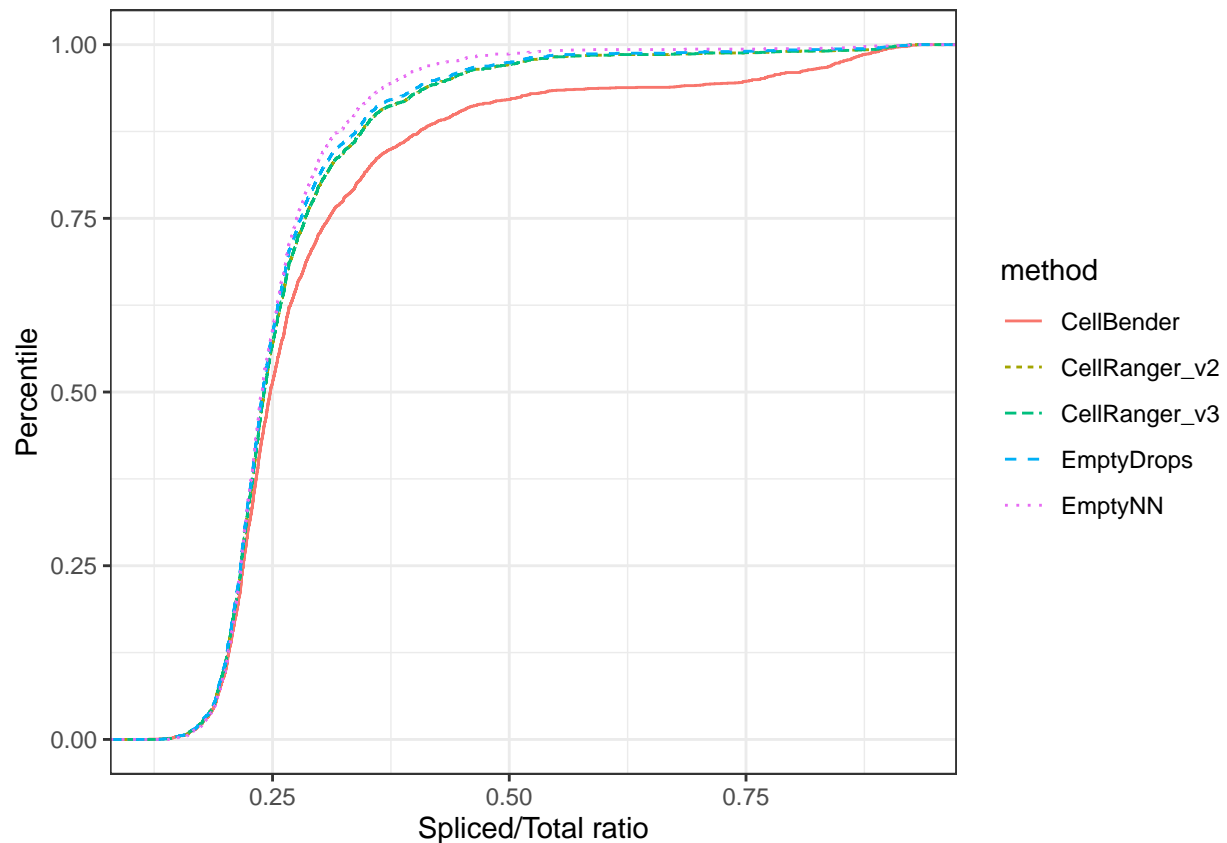


```
ggplot(tmp, aes(x=expr, colour=method,linetype=method),size=5) +
  xlab("Spliced/Total ratio")+theme_bw()+
  stat_ecdf()+ylab("Percentile")+
  coord_cartesian(xlim=c(0.25,0.5),ylim=c(0.75,1))
```

Warning: Removed 187 rows containing non-finite values (stat_ecdf).