Simulation in Different Scenario

Xinyi Lin

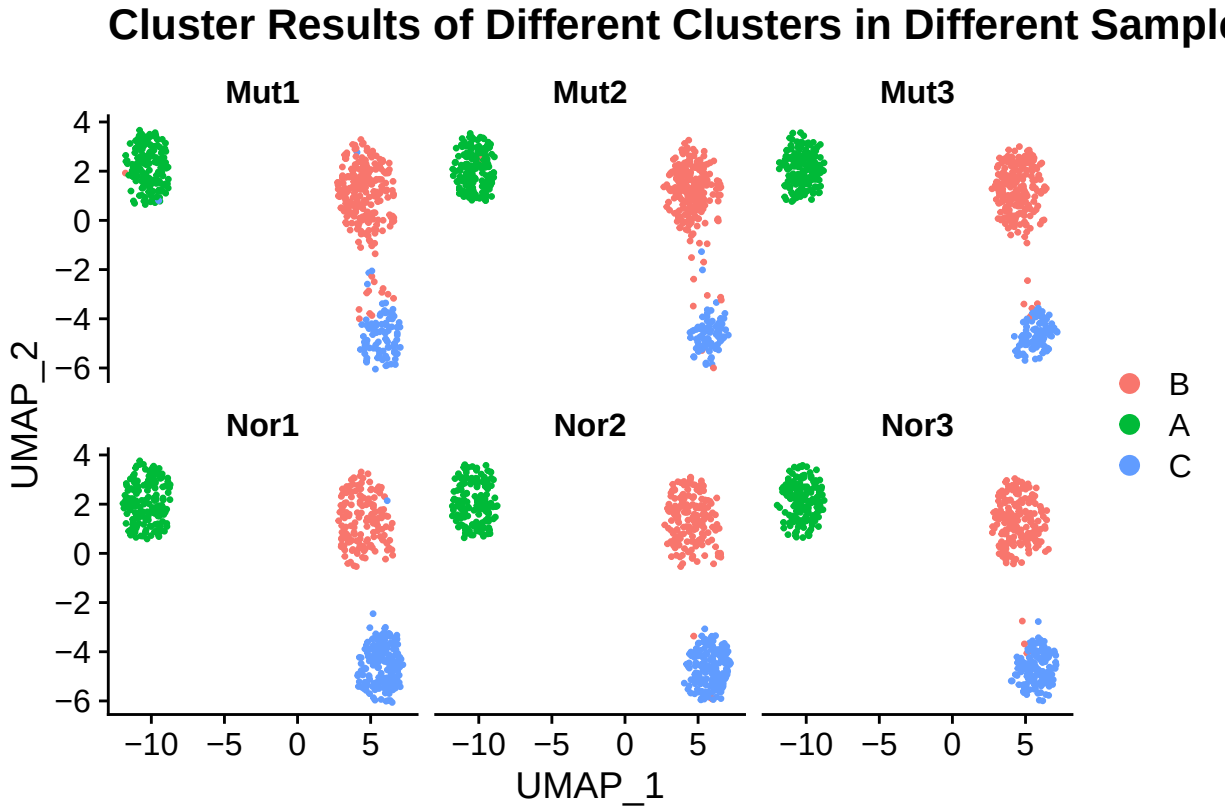
9/17/2020

Problems need to solve:

1. DCATS: label problems
2. scDC: functions don't work

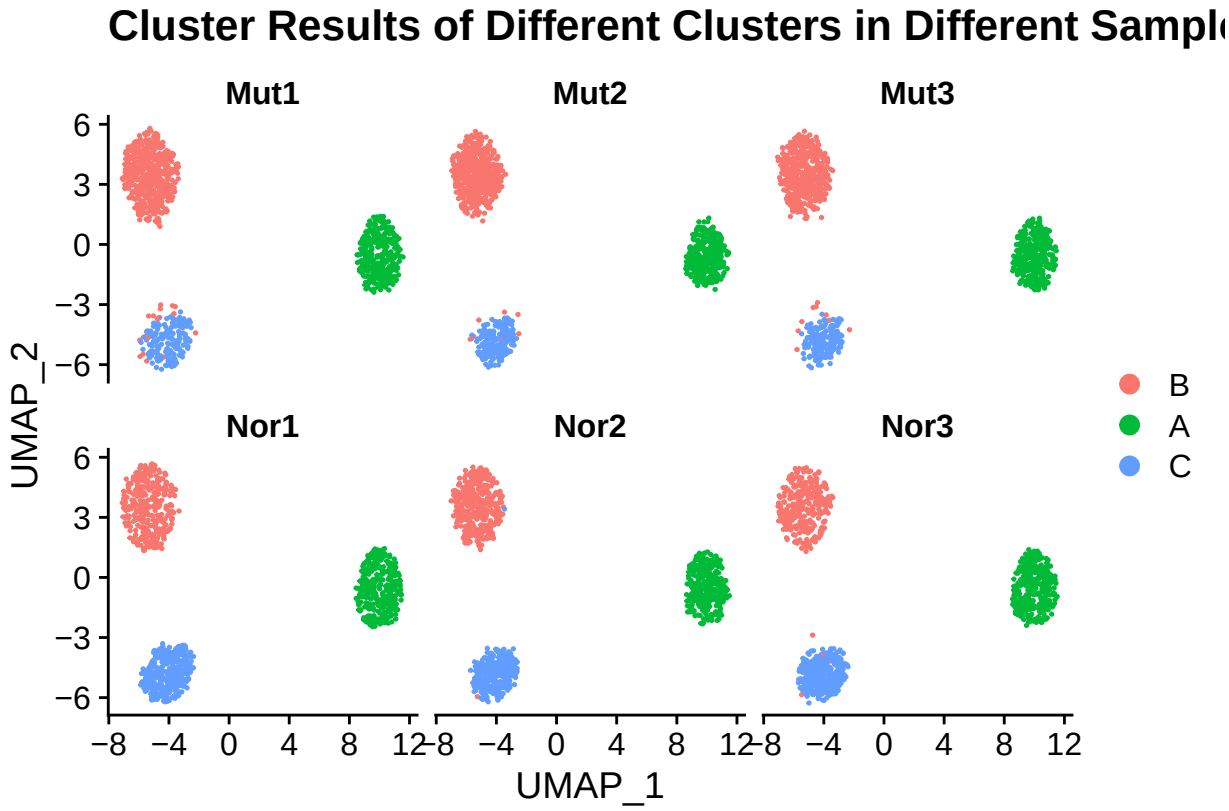
Different batch size

Batch sizes are 500, 1000, 1500, 2000.

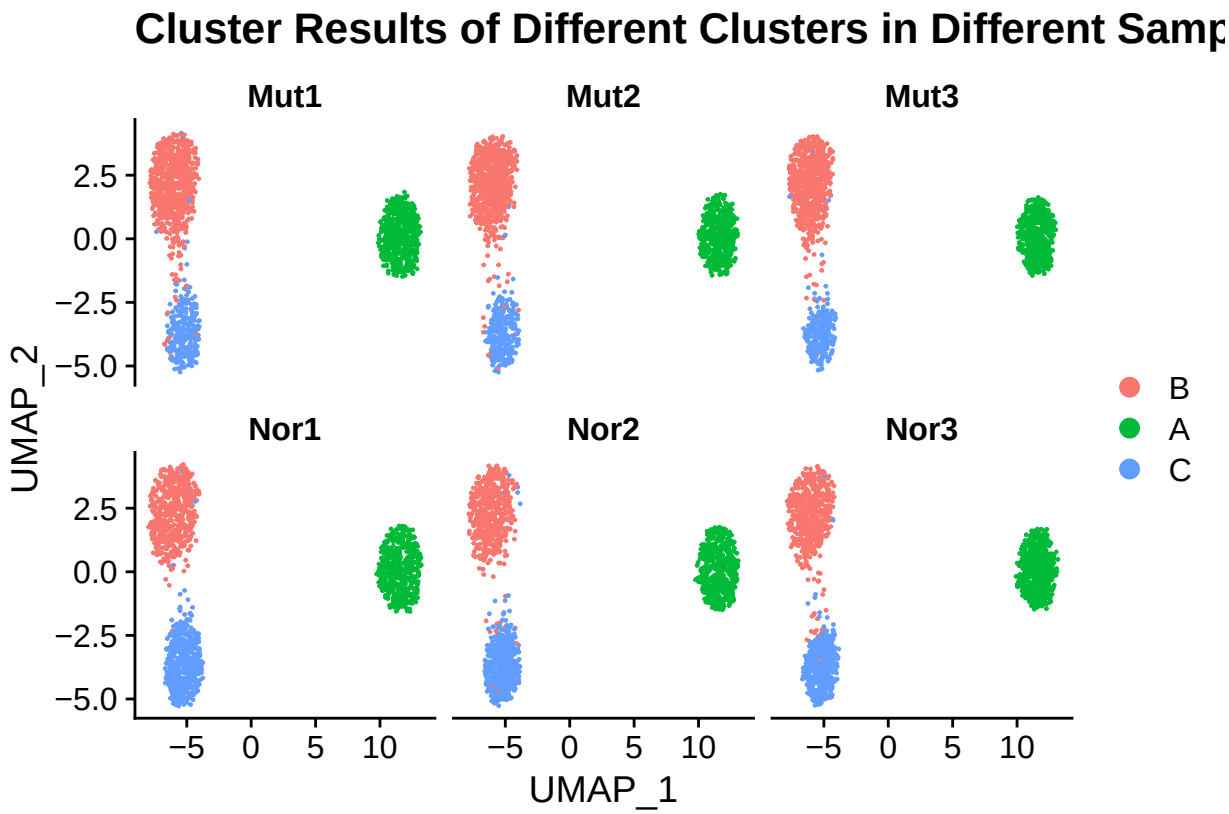



```
## $Res_df
##   cluster  dcats_pvals speckle_pvals fisher_pvals
## 1       A 6.929452e-01  5.659978e-01 4.361157e-01
```

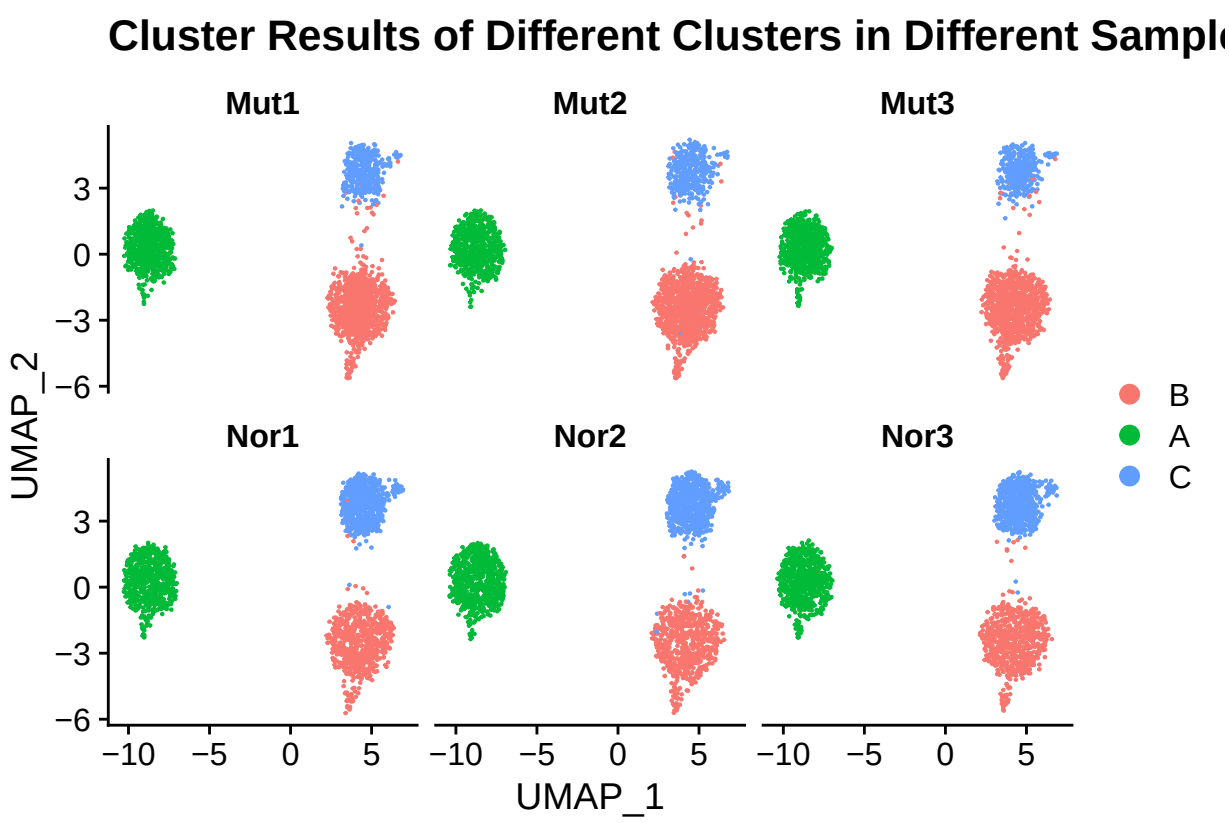
```
## 2       B 6.072355e-08  1.575240e-04 6.488365e-15
## 3       C 6.889945e-15  4.182365e-05 3.144598e-22
##
## $time_df
##   methods        time
## 1  fisher 0.007979155
## 2 sepckle 0.010969877
## 3   dcats 0.238394976
```




```
## $Res_df
##   cluster  dcats_pvals speckle_pvals fisher_pvals
## 1       A 7.742517e-01  7.417927e-01 6.795236e-01
## 2       B 2.115282e-18  2.939822e-07 2.905624e-42
## 3       C 1.027788e-41  2.022382e-07 9.059806e-51
##
## $time_df
##   methods        time
## 1  fisher 0.012964964
## 2 sepckle 0.004987001
## 3   dcats 0.260983944
```




```
## $Res_df
##   cluster  dcats_pvals speckle_pvals fisher_pvals
## 1       A 7.457880e-01  5.578942e-01 5.606287e-01
## 2       B 4.352388e-08  1.224444e-08 8.788459e-45
## 3       C 2.210894e-44  8.998110e-09 6.719757e-53
##
## $time_df
##   methods        time
## 1  fisher 0.015971899
## 2 sepckle 0.005978107
## 3   dcats 0.260437012
```

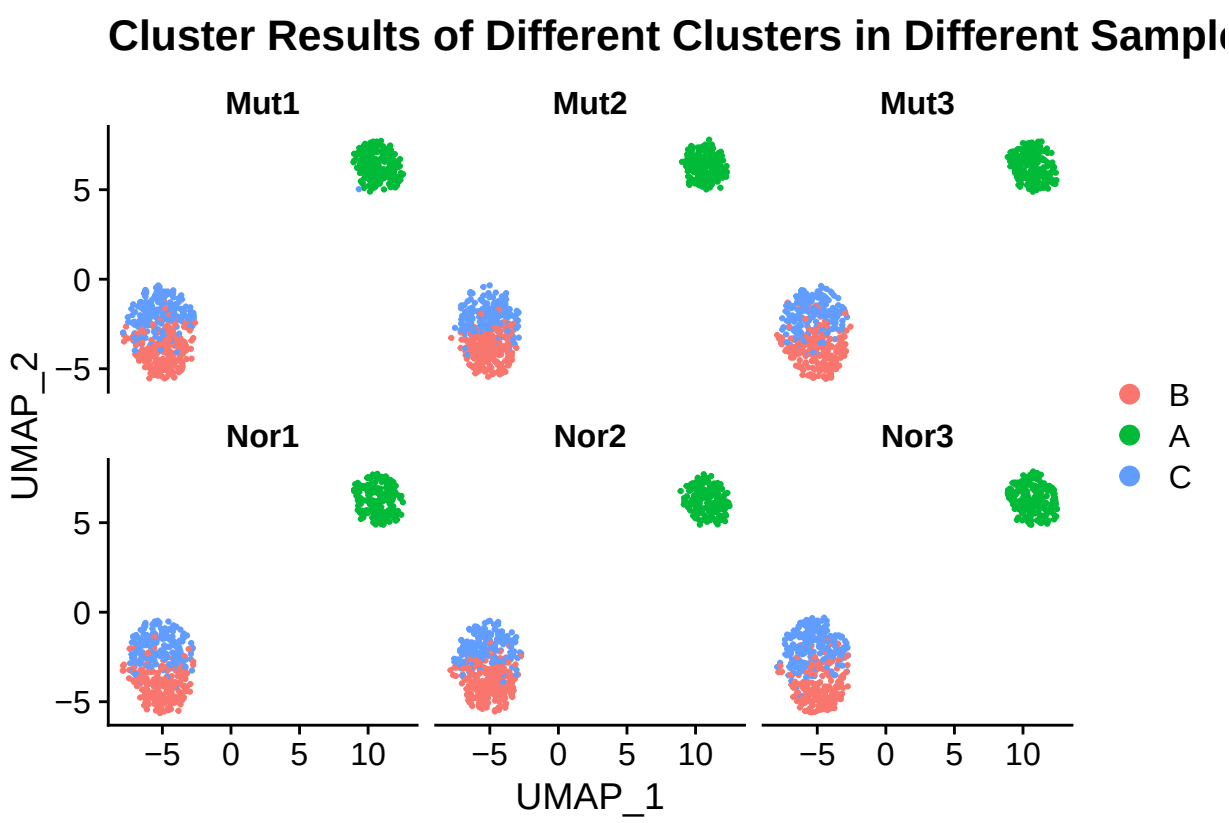



```
## $Res_df
##   cluster  dcats_pvals speckle_pvals fisher_pvals
## 1       A 7.685354e-01  6.601008e-01 6.422941e-01
## 2       B 2.679213e-18  2.057987e-09 2.325315e-68
## 3       C 1.511594e-67  1.180343e-09 5.211656e-84
##
## $time_df
##   methods        time
## 1  fisher 0.020992041
## 2 sepckle 0.006932974
## 3   dcats 0.250893831
```

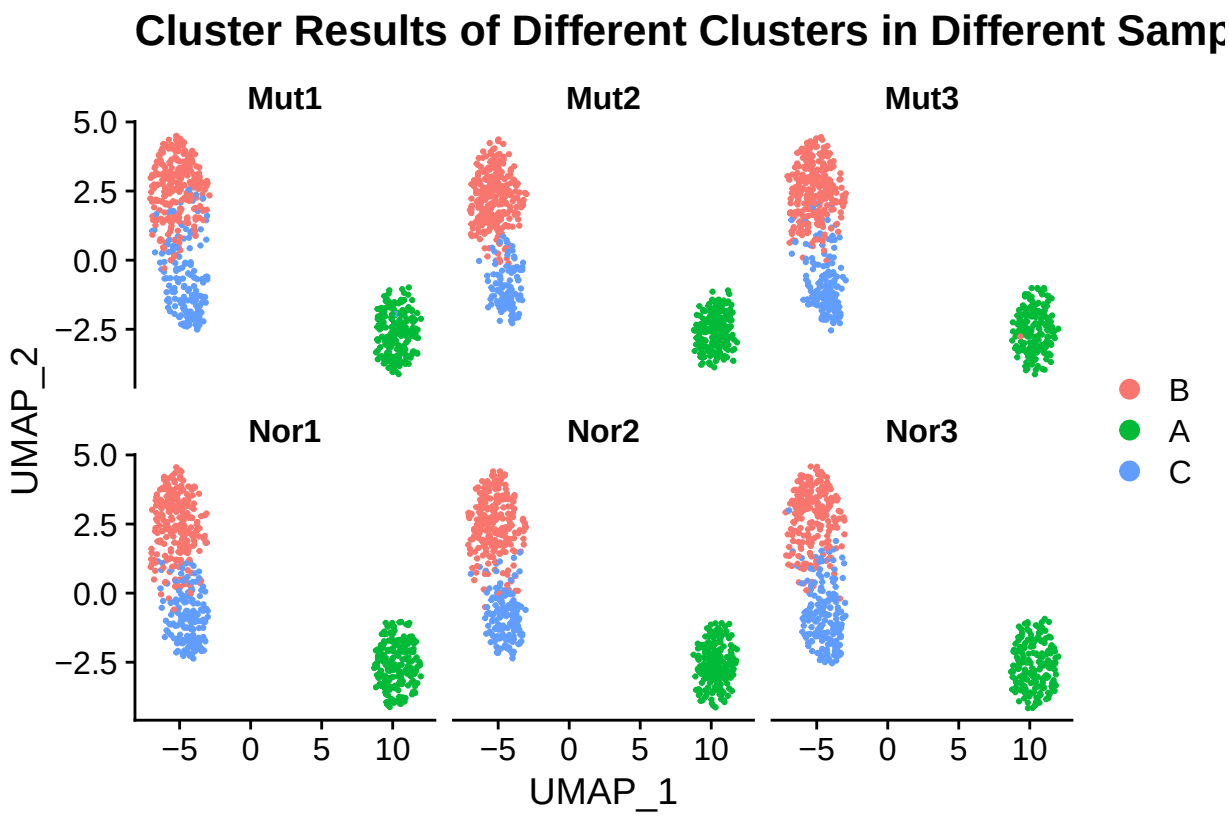
Different de.prob

The de.prob of cluster B, C are 0.02, 0.04, 0.06, 0.08.

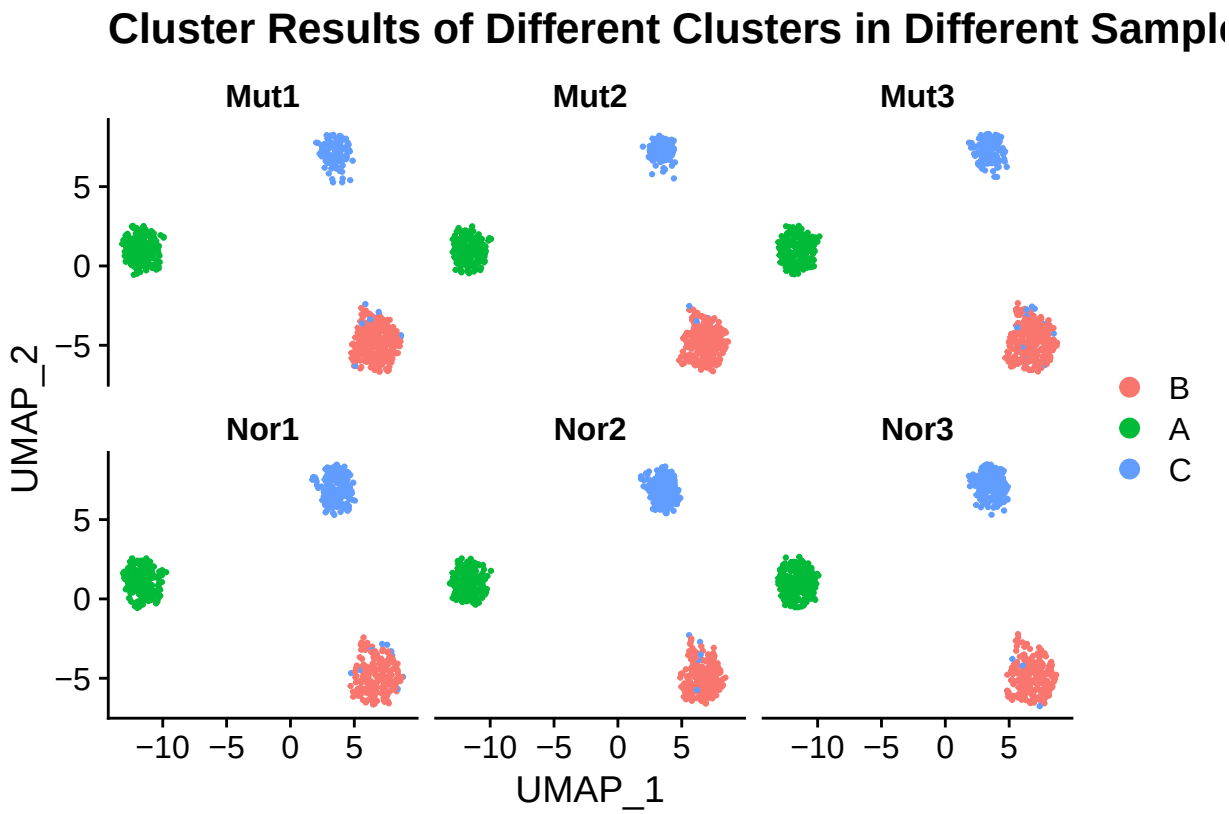
When de.prob equals to 0.04, only the fisher's exact test give the "correct" result.



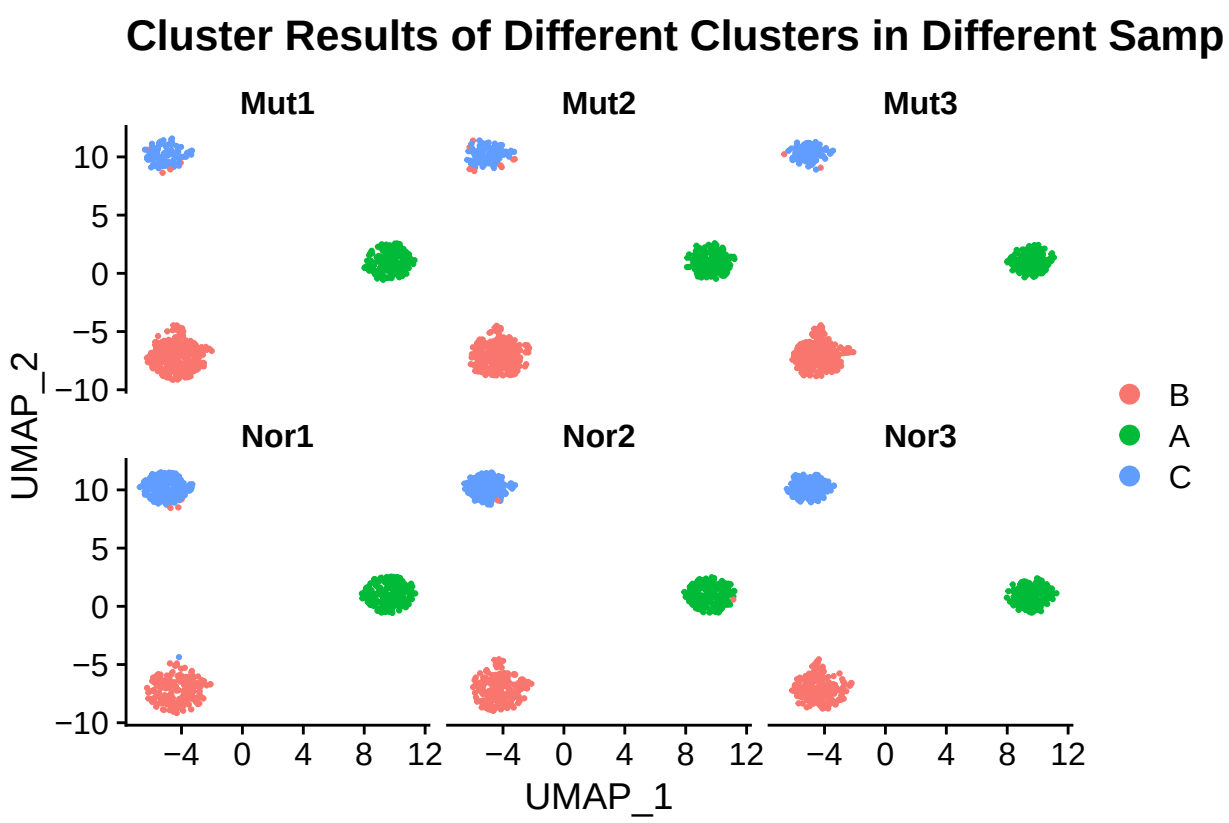

```
## $Res_df
##   cluster dcats_pvals speckle_pvals fisher_pvals
## 1       A   0.9431431     0.8301337    0.4986057
## 2       B   0.9069592     0.8301337    0.5312071
## 3       C   0.5086470     0.9781097    1.0000000
##
## $time_df
##   methods        time
## 1  fisher 0.008976221
## 2 sepckle 0.007015944
## 3   dcats 0.251144886
```




```
## $Res_df
##   cluster dcats_pvals speckle_pvals fisher_pvals
## 1       A  0.94563519     0.4400677 0.4986056694
## 2       B  0.31698036     0.1261368 0.0084599195
## 3       C  0.00768676     0.1261368 0.0001759251
##
## $time_df
##   methods        time
## 1  fisher 0.010958910
## 2 sepckle 0.004955053
## 3   dcats 0.255873919
```

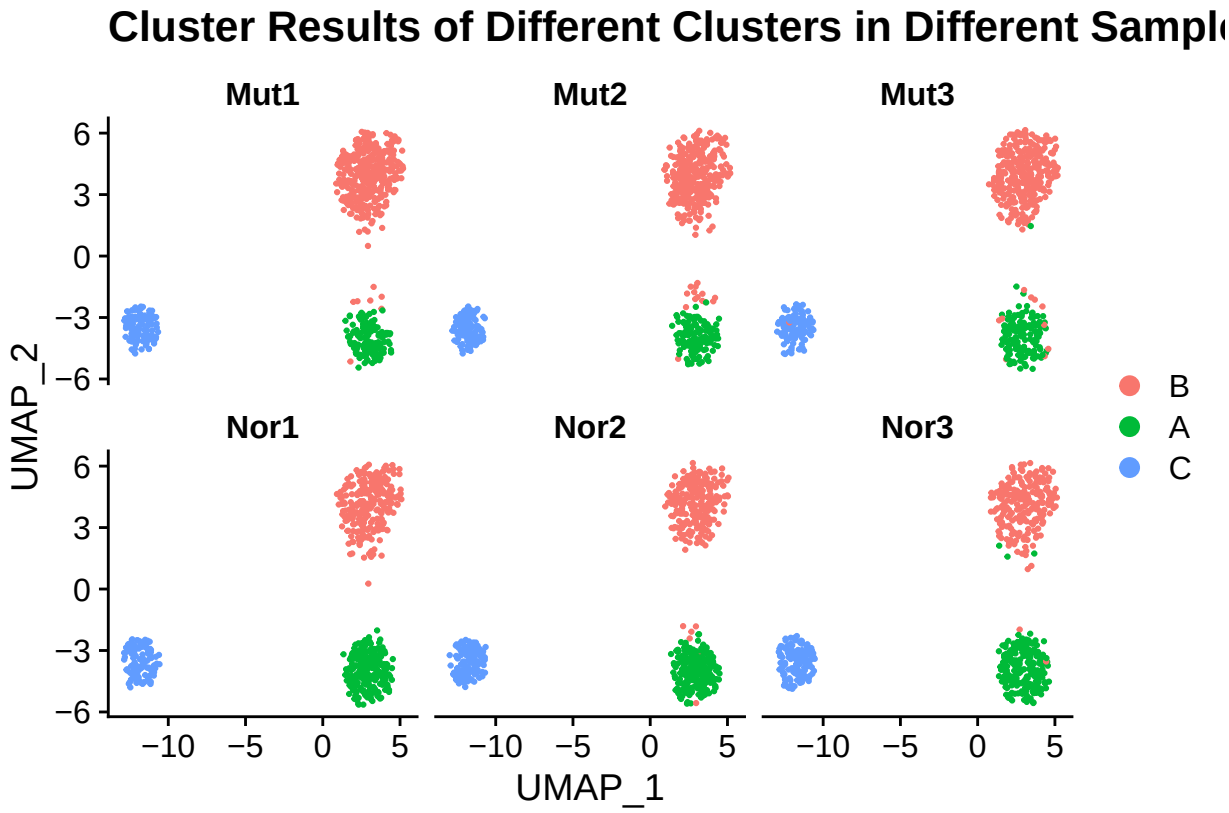



```
## $Res_df
##   cluster  dcats_pvals speckle_pvals fisher_pvals
## 1       A 8.202455e-01  5.041440e-01 4.986057e-01
## 2       B 5.978434e-06  8.724020e-06 4.010225e-15
## 3       C 4.142148e-15  2.009836e-06 5.921696e-22
##
## $time_df
##   methods        time
## 1  fisher 0.009046078
## 2 sepckle 0.004921913
## 3   dcats 0.257122040
```

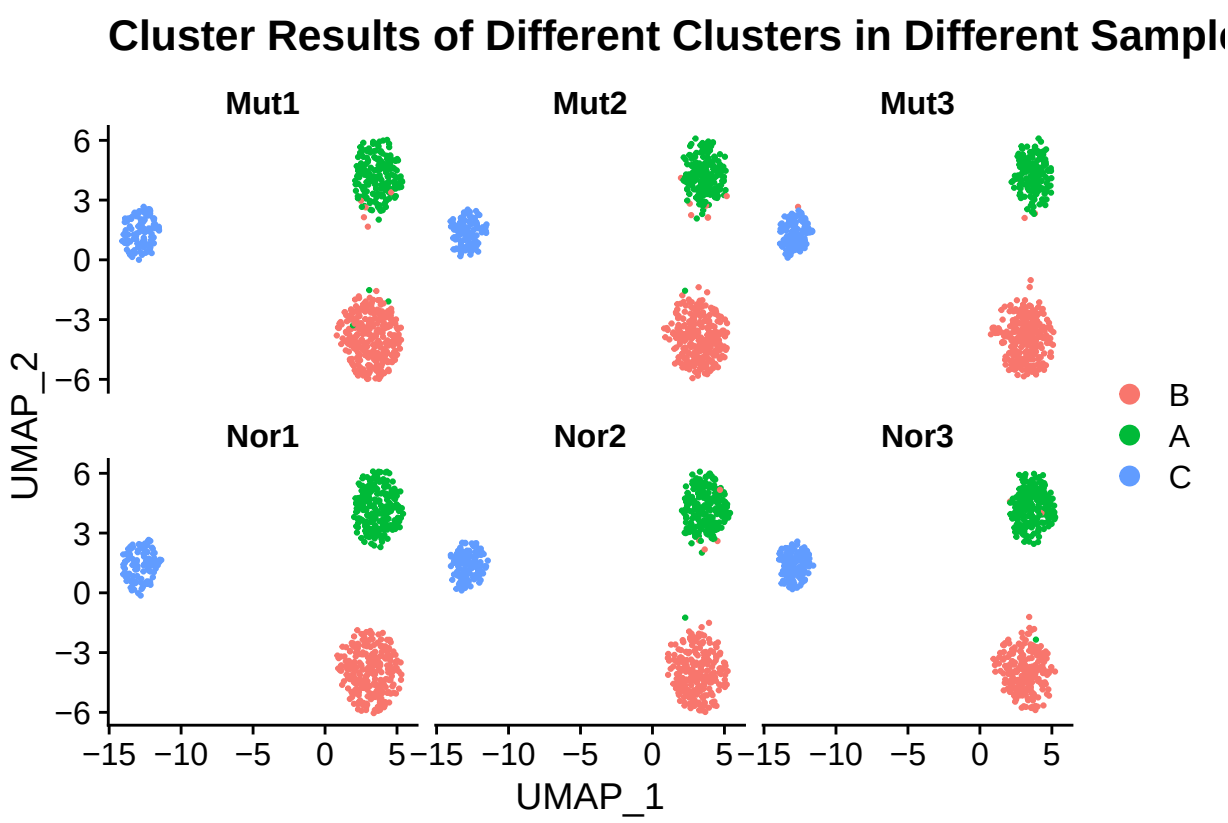



```
## $Res_df
##   cluster  dcats_pvals speckle_pvals fisher_pvals
## 1       A 6.758255e-01  5.158060e-01 4.986057e-01
## 2       B 3.188108e-12  3.284386e-06 1.383208e-20
## 3       C 1.808849e-20  6.352862e-07 6.146485e-30
##
## $time_df
##   methods        time
## 1  fisher 0.011127234
## 2 sepckle 0.005973816
## 3   dcats 0.242165089
```

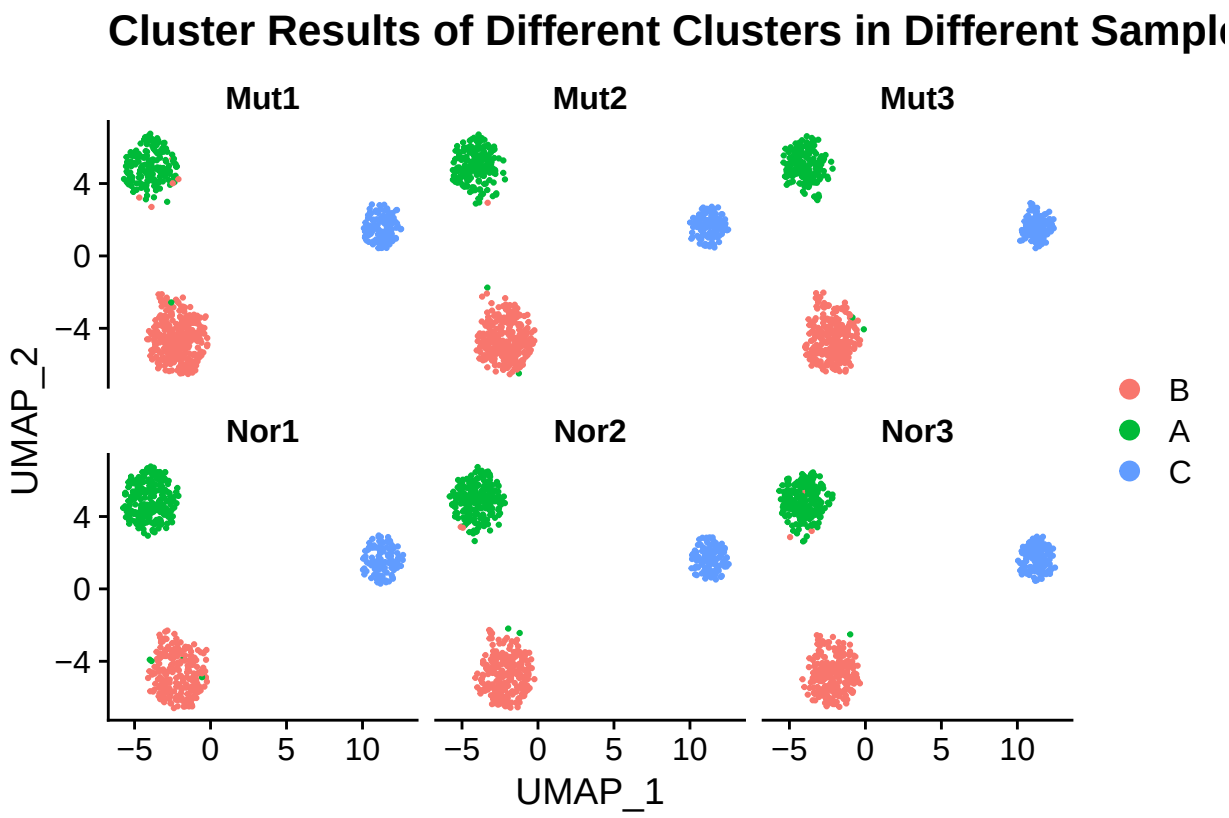
Different composition



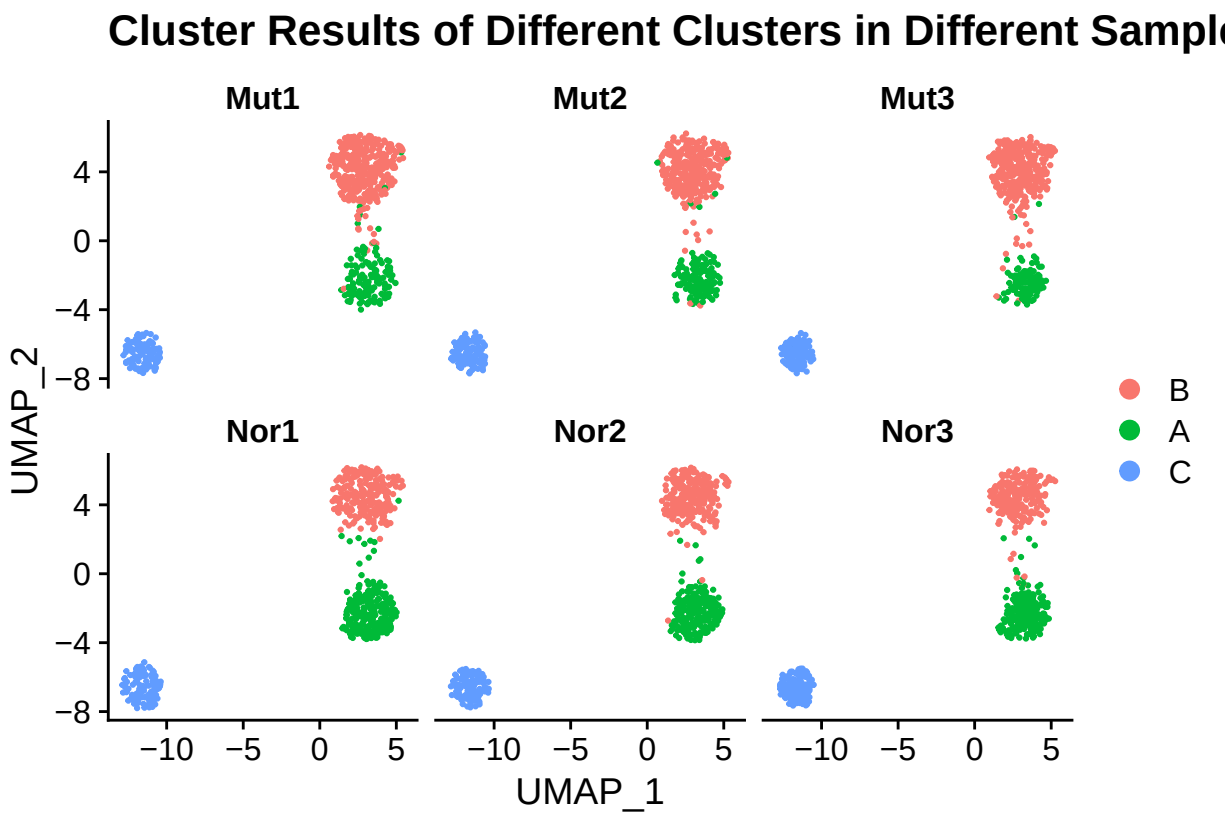

```
## $Res_df
##   cluster  dcats_pvals speckle_pvals fisher_pvals
## 1       A 3.329999e-06  4.758358e-06 3.541373e-23
## 2       B 7.174435e-06  4.758358e-06 6.055524e-24
## 3       C 2.272624e-01  3.463323e-01 2.435776e-01
##
## $time_df
##   methods        time
## 1  fisher 0.007977962
## 2 sepckle 0.004959106
## 3   dcats 0.248333931
```




```
## $Res_df
##   cluster dcats_pvals speckle_pvals fisher_pvals
## 1       A  0.01543553  0.0005798027 4.920662e-07
## 2       B  0.01113654  0.0005798027 1.641263e-06
## 3       C  1.00000000  1.0000000000 1.000000e+00
##
## $time_df
##   methods        time
## 1  fisher 0.007974863
## 2 sepckle 0.003992081
## 3   dcats 0.247341871
```




```
## $Res_df
##   cluster dcats_pvals speckle_pvals fisher_pvals
## 1       A 0.002236473   0.000165966 1.874273e-10
## 2       B 0.003602618   0.000165966 9.801246e-10
## 3       C 1.000000000   1.000000000 1.000000e+00
##
## $time_df
##   methods        time
## 1  fisher 0.007977962
## 2 sepckle 0.003988981
## 3   dcats 0.245347023
```




```
## $Res_df
##   cluster  dcats_pvals speckle_pvals fisher_pvals
## 1       A 3.410972e-06  1.041730e-06 2.859680e-30
## 2       B 1.502514e-05  1.283716e-06 6.121191e-26
## 3       C 1.000000e+00  1.000000e+00 1.000000e+00
##
## $time_df
##   methods        time
## 1  fisher 0.006981134
## 2 sepckle 0.004987955
## 3   dcats 0.252317905
```

Codes

```
knitr::opts_chunk$set(echo = FALSE)
library(splatter)
library(Seurat)
library(speckle)
library(DCATS)
library(ggplot2)
library(tidyverse)
source("functions.r")
set.seed(123)
probNor = c(1/3,1/3,1/3)
```

```
probMut = c(1/2,1/6,1/3)
de_prob = c(0.1,0.1,0.5)

batch_size_list = seq(500,2000,500)
setresolu_list = c(0.5, 0.3, 0.3, 0.3)
for (i in 1:4) {
  batch_size = batch_size_list[i]
  setresolu = setresolu_list[i]

  sim_list = simualtion(probNor, probMut, resolution, de_prob, batch_size)
  integratedSamples = runSeurat(sim_list, batch_size)
  time = rep(NA,3)
  plot = DimPlot(integratedSamples, ncol = 3, reduction = "umap", split.by = "batch") + ggtitle("Cluste
  print(plot)
  dfRes = data.frame(clusterRes = integratedSamples@active.ident, batch = integratedSamples$batch, condi
    tibble::rownames_to_column("cellID")
  ## Fisher's exact test
  dfCount = dfRes %>%
    group_by(condition, clusterRes) %>%
    summarise(n = n()) %>%
    pivot_wider(names_from = "clusterRes", values_from = "n") %>%
    mutate(nonA = B + C,
           nonB = A + C,
           nonC = A + B) %>%
    select(A, B, C:nonC)
  t1start = Sys.time()
  fisher_pvals = rep(NA,3)
  for (i in 1:3){
    fisher_pvals[i] = fisher.test(dfCount[,c(i+1,i+4)])$p.value
    }
  time[1] = Sys.time() - t1start
  ## speckle
  t2start = Sys.time()
  speckleRes = propeller(clusters = dfRes$clusterRes, sample = dfRes$batch, group = dfRes$condition)
  time[2] = Sys.time() - t2start
  #print(speckleRes)
  speckleP = data.frame(cluster = speckleRes$BaselineProp.clusters, speckle_pvals = speckleRes$FDR)
  ## DCATS
  celllabels_orig = sim_list$origLabels
  conf.mat<-table(Idents(integratedSamples), celllabels_orig)
  true.conf<-t(t(conf.mat)/apply(conf.mat,2,sum))
  condition = integratedSamples@meta.data$condition
  condNor<-Idents(integratedSamples)[condition == "Normal"]
  condMut<-Idents(integratedSamples)[condition == "Mutate"]
  countNor = table(sim_list$batchNor, relevel(condNor, "A"))
  countMut = table(sim_list$batchMut, relevel(condMut, "A"))
  t3start = Sys.time()
  dcatsRes = dcats_fit(countNor, countMut, true.conf)
  time[3] = Sys.time() - t3start
  #print(dcatsRes)
  dcatsP = data.frame(cluster = rownames(dcatsRes), dcats_pvals = dcatsRes$pvals)
  Res_df = merge(dcatsP, speckleP, by = "cluster") %>%
    mutate(fisher_pvals = fisher_pvals)
```

```
  time_df = data.frame(methods = c("fisher", "sepckle", "dcats"), time = time)
  print(list(Res_df = Res_df, time_df = time_df))
}
set.seed(123)
probNor = c(1/3,1/3,1/3)
probMut = c(1/2,1/6,1/3)
batch_size = 600

de_prob_list = seq(0.02, 0.08, 0.02)
setresolu_list = c(0.4, 0.4, 0.4, 0.3)
for (i in 1:4) {
  de_prob = c(de_prob_list[i], de_prob_list[i], 0.5)
  setresolu = setresolu_list[i]

  sim_list = simualtion(probNor, probMut, resolution, de_prob, batch_size)
  integratedSamples = runSeurat(sim_list, batch_size)
  time = rep(NA,3)
  plot = DimPlot(integratedSamples, ncol = 3, reduction = "umap", split.by = "batch") + ggtitle("Cluste
  print(plot)
  dfRes = data.frame(clusterRes = integratedSamples@active.ident, batch = integratedSamples$batch, condi
    tibble::rownames_to_column("cellID")
  ## Fisher's exact test
  dfCount = dfRes %>%
    group_by(condition, clusterRes) %>%
    summarise(n = n()) %>%
    pivot_wider(names_from = "clusterRes", values_from = "n") %>%
    mutate(nonA = B + C,
           nonB = A + C,
           nonC = A + B) %>%
    select(A, B, C:nonC)
  t1start = Sys.time()
  fisher_pvals = rep(NA,3)
  for (i in 1:3){
    fisher_pvals[i] = fisher.test(dfCount[,c(i+1,i+4)])$p.value
    }
  time[1] = Sys.time() - t1start
  ## speckle
  t2start = Sys.time()
  speckleRes = propeller(clusters = dfRes$clusterRes, sample = dfRes$batch, group = dfRes$condition)
  time[2] = Sys.time() - t2start
  #print(speckleRes)
  speckleP = data.frame(cluster = speckleRes$BaselineProp.clusters, speckle_pvals = speckleRes$FDR)
  ## DCATS
  celllabels_orig = sim_list$origLabels
  conf.mat<-table(Idents(integratedSamples), celllabels_orig)
  true.conf<-t(t(conf.mat)/apply(conf.mat,2,sum))
  condition = integratedSamples@meta.data$condition
  condNor<-Idents(integratedSamples)[condition == "Normal"]
  condMut<-Idents(integratedSamples)[condition == "Mutate"]
  countNor = table(sim_list$batchNor, relevel(condNor, "A"))
  countMut = table(sim_list$batchMut, relevel(condMut, "A"))
  t3start = Sys.time()
  dcatsRes = dcats_fit(countNor, countMut, true.conf)
```

```
  time[3] = Sys.time() - t3start
  #print(dcatsRes)
  dcatsP = data.frame(cluster = rownames(dcatsRes), dcats_pvals = dcatsRes$pvals)
  Res_df = merge(dcatsP, speckleP, by = "cluster") %>%
    mutate(fisher_pvals = fisher_pvals)
  time_df = data.frame(methods = c("fisher", "sepckle", "dcats"), time = time)
  print(list(Res_df = Res_df, time_df = time_df))
}
set.seed(123)
probNor = c(0.4, 0.4, 0.2)
batch_size = 600
de_prob = c(0.06,0.06,0.5)

probMut_list = c(0.2, 0.3, 0.5, 0.6)
setresolu_list = c(0.3, 0.3, 0.3, 0.3)
for (i in 1:4) {
  probMut = c(probMut_list[i], 0.8 - probMut_list[i], 0.2)
  setresolu = setresolu_list[i]

  sim_list = simualtion(probNor, probMut, resolution, de_prob, batch_size)
  integratedSamples = runSeurat(sim_list, batch_size)
  time = rep(NA,3)
  plot = DimPlot(integratedSamples, ncol = 3, reduction = "umap", split.by = "batch") + ggtitle("Cluste
  print(plot)
  dfRes = data.frame(clusterRes = integratedSamples@active.ident, batch = integratedSamples$batch, condi
    tibble::rownames_to_column("cellID")
  ## Fisher's exact test
  dfCount = dfRes %>%
    group_by(condition, clusterRes) %>%
    summarise(n = n()) %>%
    pivot_wider(names_from = "clusterRes", values_from = "n") %>%
    mutate(nonA = B + C,
           nonB = A + C,
           nonC = A + B) %>%
    select(A, B, C:nonC)
  t1start = Sys.time()
  fisher_pvals = rep(NA,3)
  for (i in 1:3){
    fisher_pvals[i] = fisher.test(dfCount[,c(i+1,i+4)])$p.value
    }
  time[1] = Sys.time() - t1start
  ## speckle
  t2start = Sys.time()
  speckleRes = propeller(clusters = dfRes$clusterRes, sample = dfRes$batch, group = dfRes$condition)
  time[2] = Sys.time() - t2start
  #print(speckleRes)
  speckleP = data.frame(cluster = speckleRes$BaselineProp.clusters, speckle_pvals = speckleRes$FDR)
  ## DCATS
  celllabels_orig = sim_list$origLabels
  conf.mat<-table(Idents(integratedSamples), celllabels_orig)
  true.conf<-t(t(conf.mat)/apply(conf.mat,2,sum))
  #print(true.conf)
  condition = integratedSamples@meta.data$condition
```

```
  condNor<-Idents(integratedSamples)[condition == "Normal"]
  condMut<-Idents(integratedSamples)[condition == "Mutate"]
  countNor = table(sim_list$batchNor, relevel(condNor, "A"))
  countMut = table(sim_list$batchMut, relevel(condMut, "A"))
  t3start = Sys.time()
  dcatsRes = dcats_fit(countNor, countMut, true.conf)
  time[3] = Sys.time() - t3start
  #print(dcatsRes)
  dcatsP = data.frame(cluster = rownames(dcatsRes), dcats_pvals = dcatsRes$pvals)
  Res_df = merge(dcatsP, speckleP, by = "cluster") %>%
    mutate(fisher_pvals = fisher_pvals)
  time_df = data.frame(methods = c("fisher", "sepckle", "dcats"), time = time)
  print(list(Res_df = Res_df, time_df = time_df))
}
```