Explore scATAC-Seq, 6 cell lines-2

Zhixiang Lin

2/21/2017

A subset of cells (~1,000) with large number of reads are retained.

Clustering after collapsing cells, randomly collapse 20 cells.

For each sample, pick the top peaks, calculate the percentage of overlap

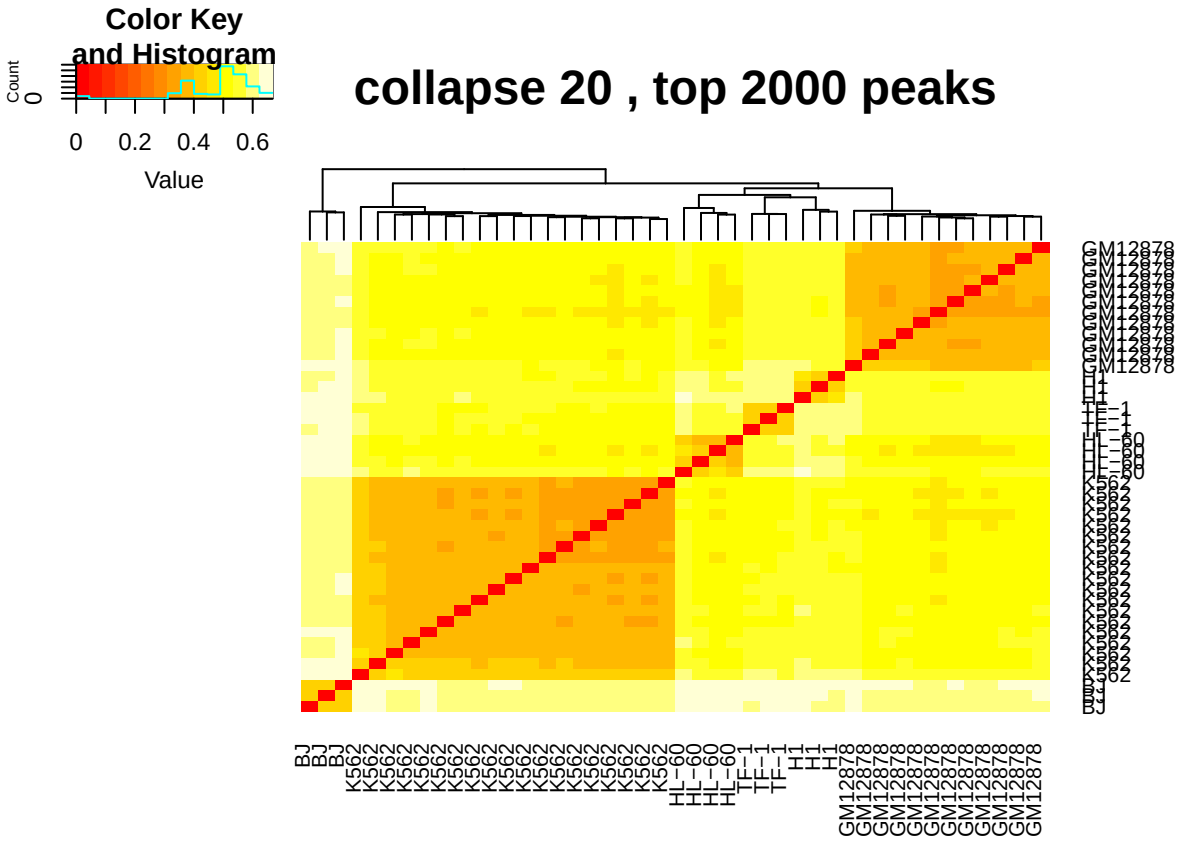
```
set.seed(123)
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

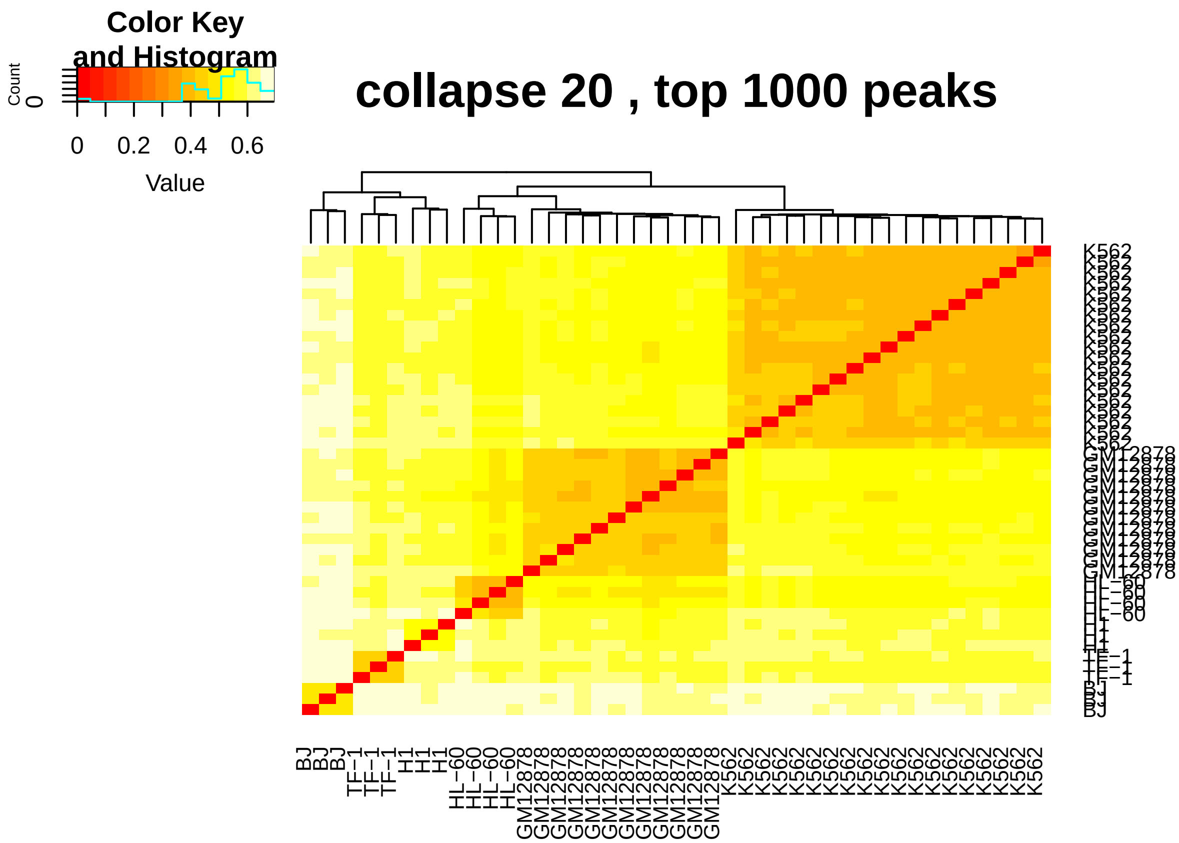
```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
caloverlap <- function(x, y, top){
  xtmp <- 0*x
  ytmp <- 0*y
  xtmp[order(x, decreasing=T)[1:top]] <- 1
  ytmp[order(y, decreasing=T)[1:top]] <- 1
  sum((xtmp+ytmp)==2)/top
}
```

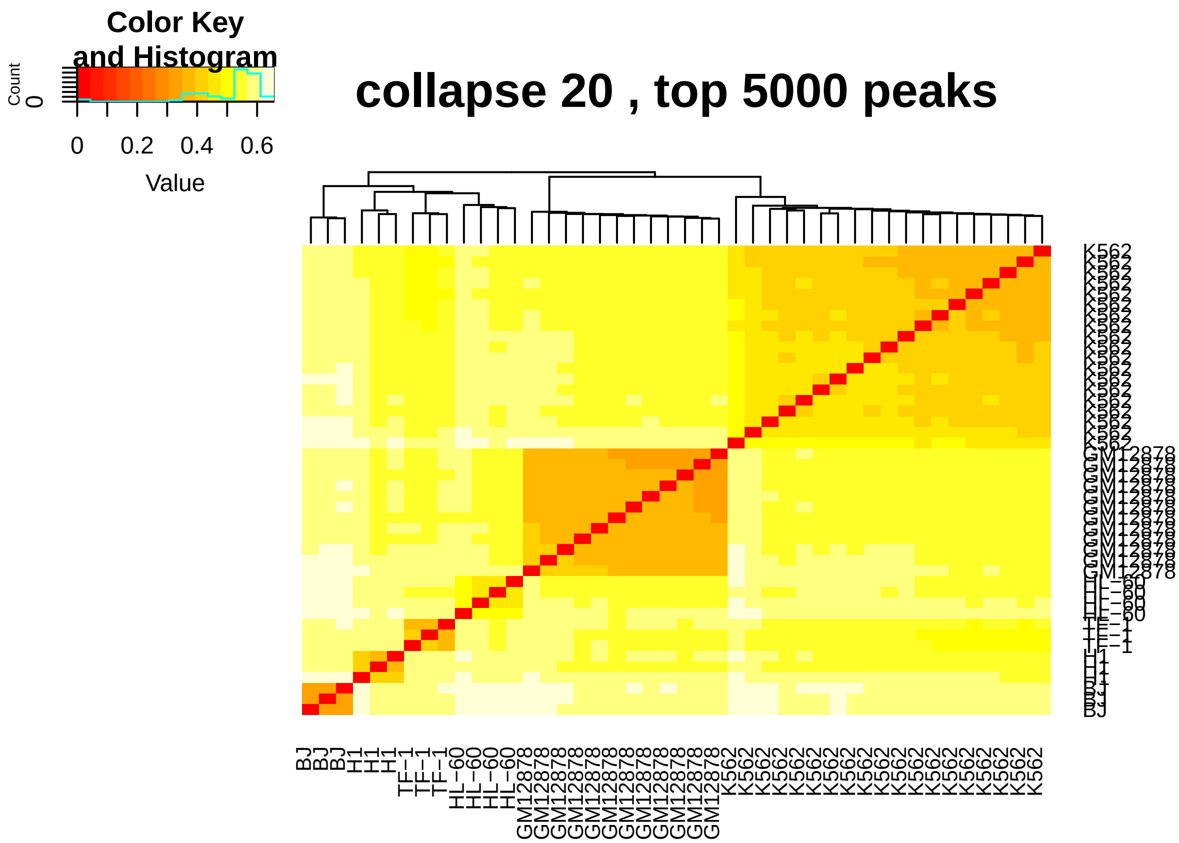
```
coll <- 20
top <- 2000
fn <- paste("/Users/zhixianglin/Documents/collaboration/mahdi/scATAC/CellLines6/dataColl",coll, ".rda",
load(fn)
data <- dataColl20$ForGroundC
cells <- dataColl20$SampleSubC
num <- ncol(data)
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top", 
```




```
top <- 1000
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top",
```



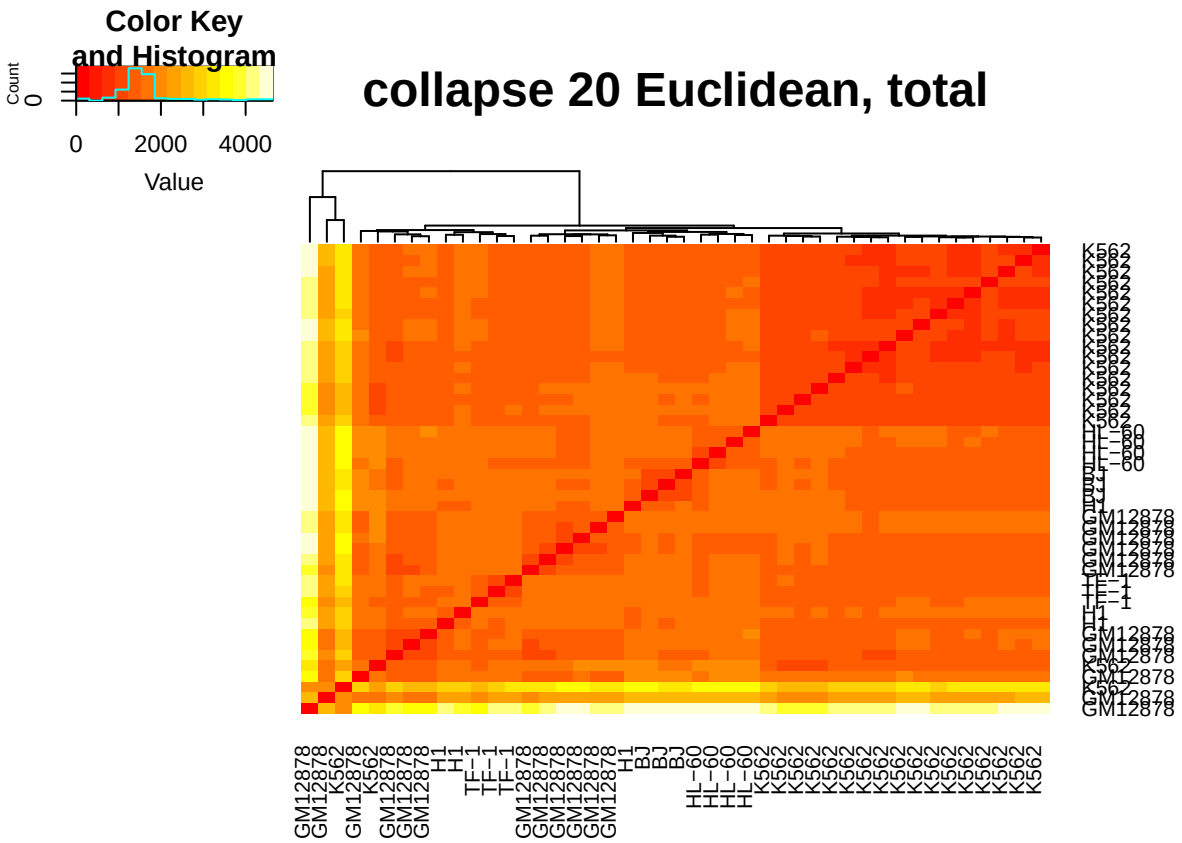

```
top <- 5000
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top", t
```



Euclidean distance after normalize by total reads in the peak

```
cs <- colSums(data)
normfactor1 <- 1/cs*median(cs)

distM <- as.matrix(dist(t(data*normfactor1)))
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, "Euclidea
```

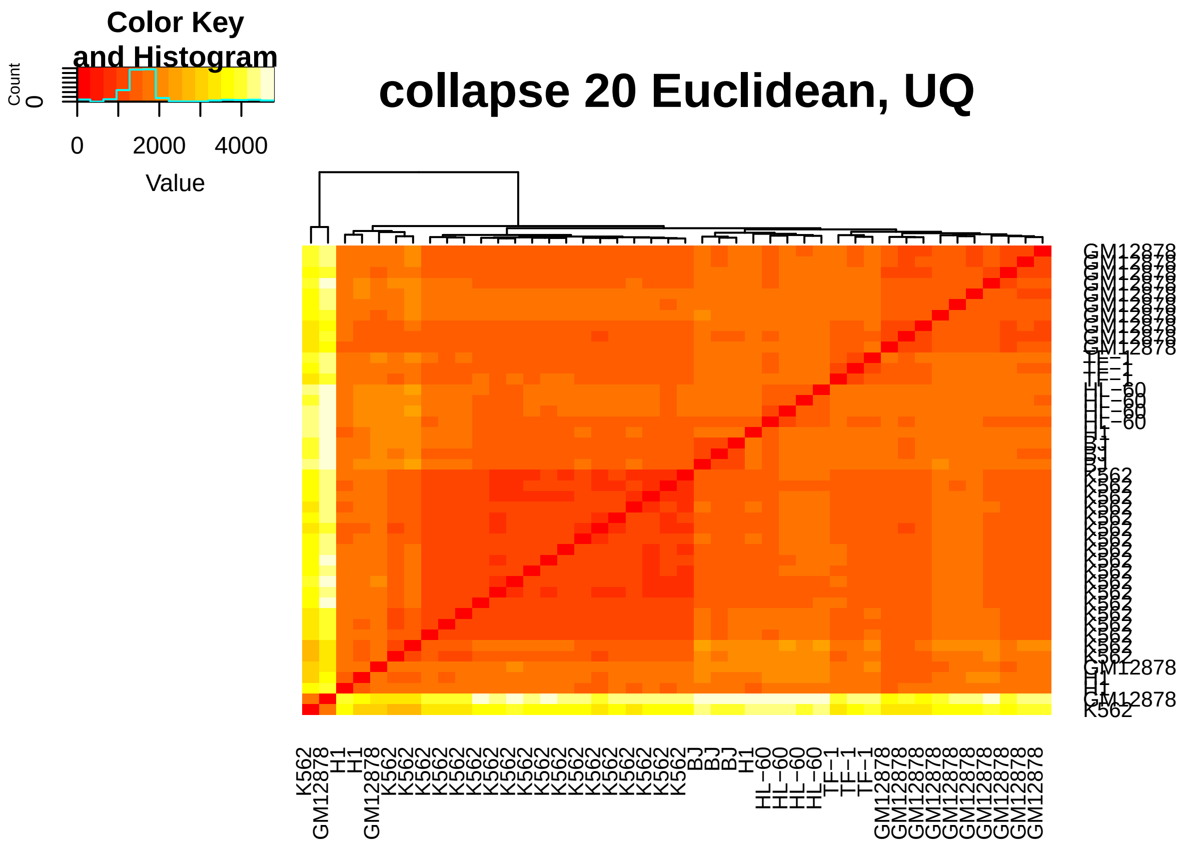


Euclidean distance after upper-quartile normalization, note that the upper-quartile is still small

```
uq <- apply(data, 2, function(x){quantile(x[x!=0], 0.75)})
print(uq)
```

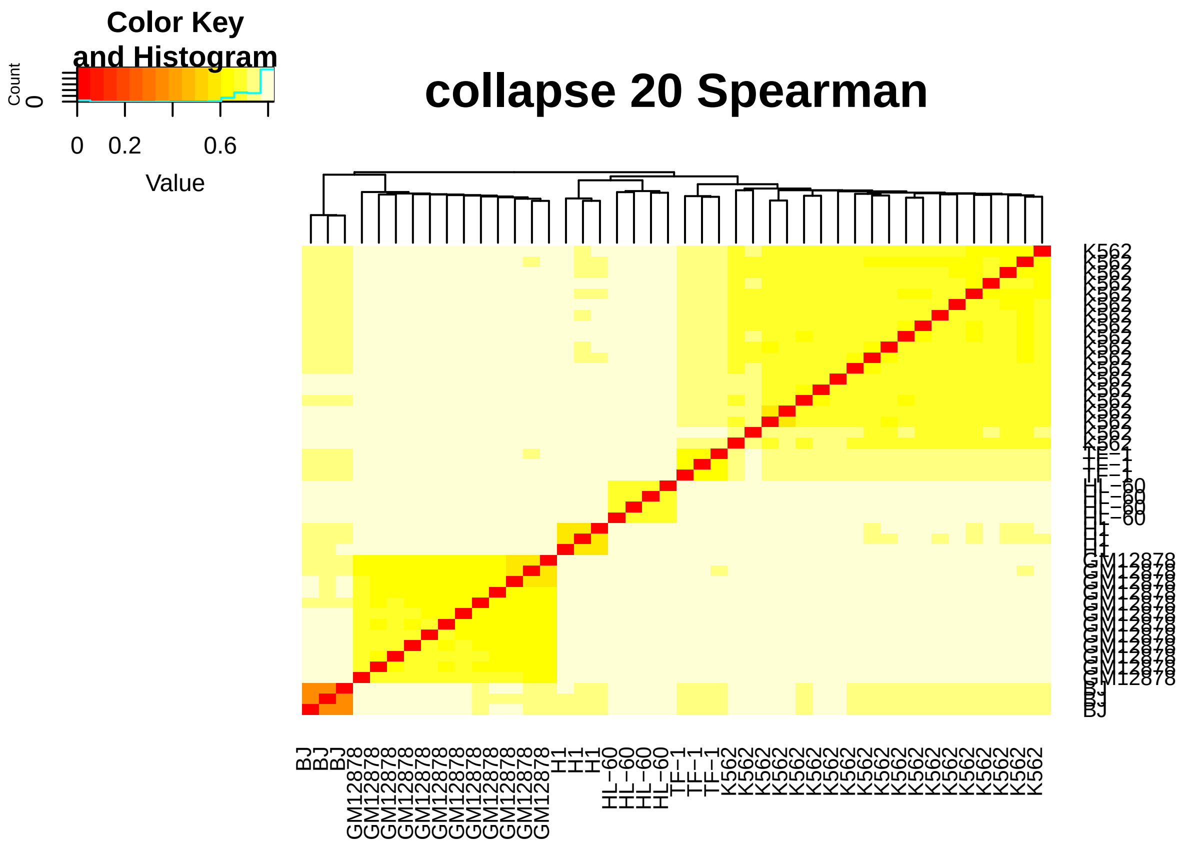
```
##  [1] 8 8 8 6 6 6 6 6 5 6 4 6 6 5 6 6 6 6 4 5 4 5 4 4 4 4 5 4 4 4 4 4 4 4 4
## [36] 4 4 4 4 4 4 5 5 5
```

```
normfactor2 <- 1/uq*median(uq)
distM <- as.matrix(dist(t(data*normfactor2)))
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, "Euclidean
```



Spearman correlation

```
datascor <- 1-cor(data, method="spearman")
colnames(datascor) <- cells
rownames(datascor) <- cells
heatmap.2(datascor, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, "Spearm
```



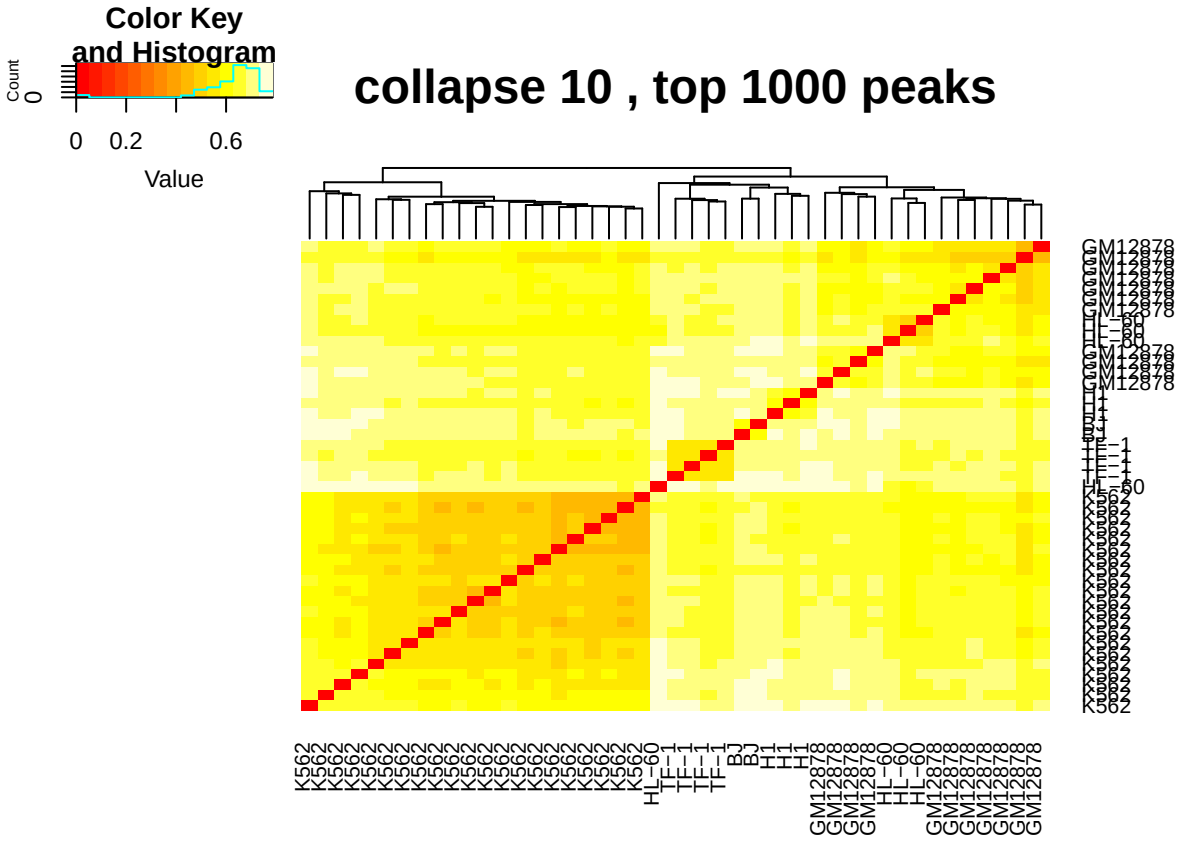
Clustering after collapsing cells, randomly collapse 10 cells.

```
coll <- 10
fn <- paste("/Users/zhixianglin/Documents/collaboration/mahdi/scATAC/CellLines6/dataColl",coll, ".rda",
load(fn)
data <- dataColl10$ForGroundC
cells <- dataColl10$SampleSubC
print(table(data[,1]))
```

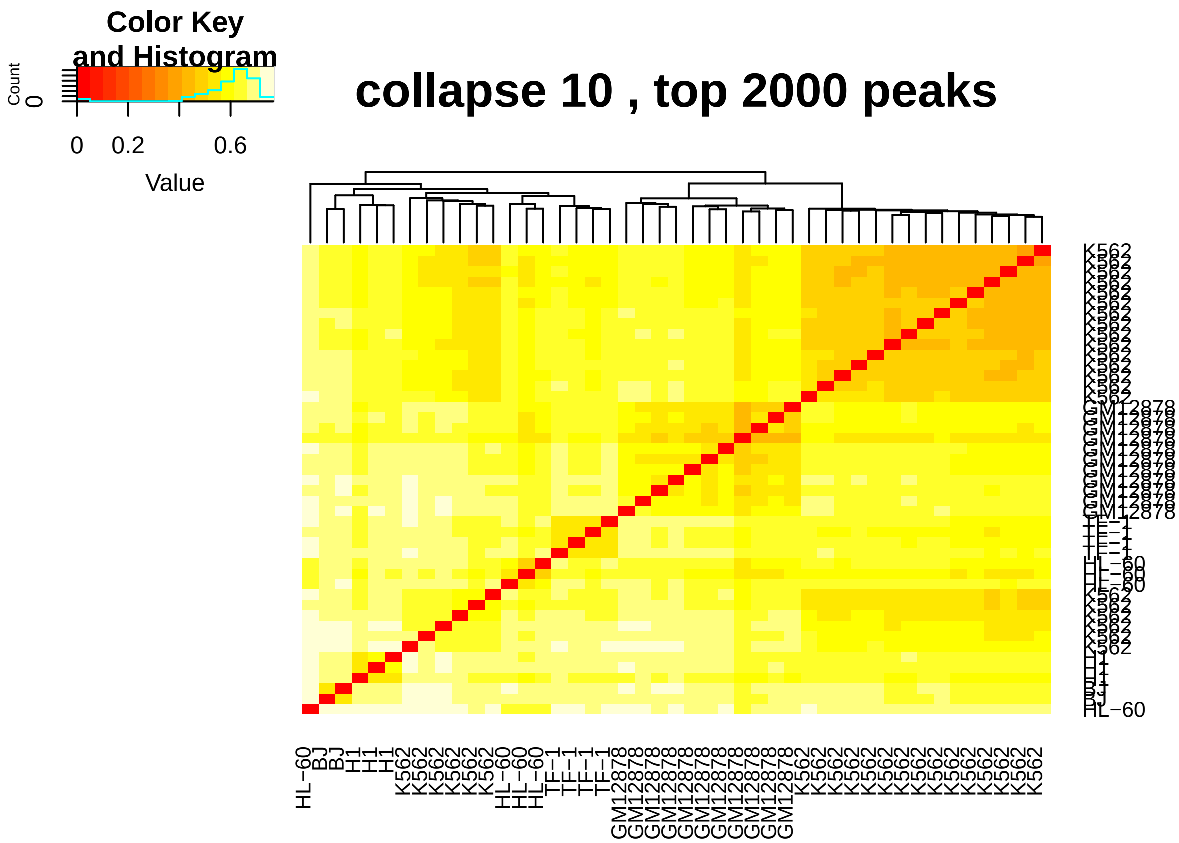
```
##
##       0      1      2      3      4      5      6      7      8      9
## 124450    793   8186    367   3445    250   1853    149   1121     93
##      10     11     12     13     14     15     16     17     18     19
##     711     60    434     56    281     37    174     33    129     18
##      20     21     22     23     24     25     26     27     28     29
##      66     12     34      6     25      3     27      2      8      5
##      30     31     32     34     36     38     42     46     51    172
##       7      2      4      2      1      1      1      1      1      1
```

```
sub <- sample(ncol(data), round(ncol(data)/2))
data <- data[,sub]
cells <- cells[sub]
num <- ncol(data)
```

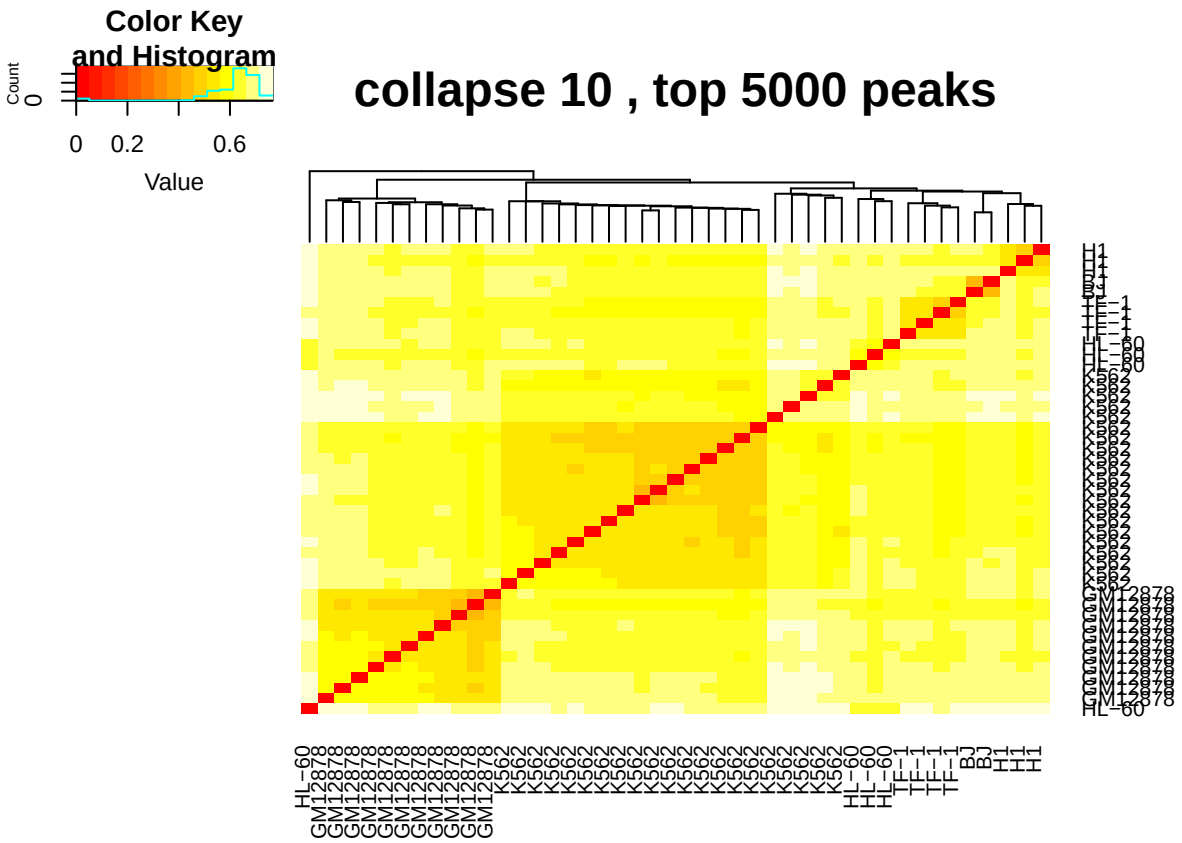
```
top <- 1000
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top", 
```




```
top <- 2000
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top", 
```




```
top <- 5000
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top", 
```



Clustering after collapsing cells, randomly collapse 5 cells.

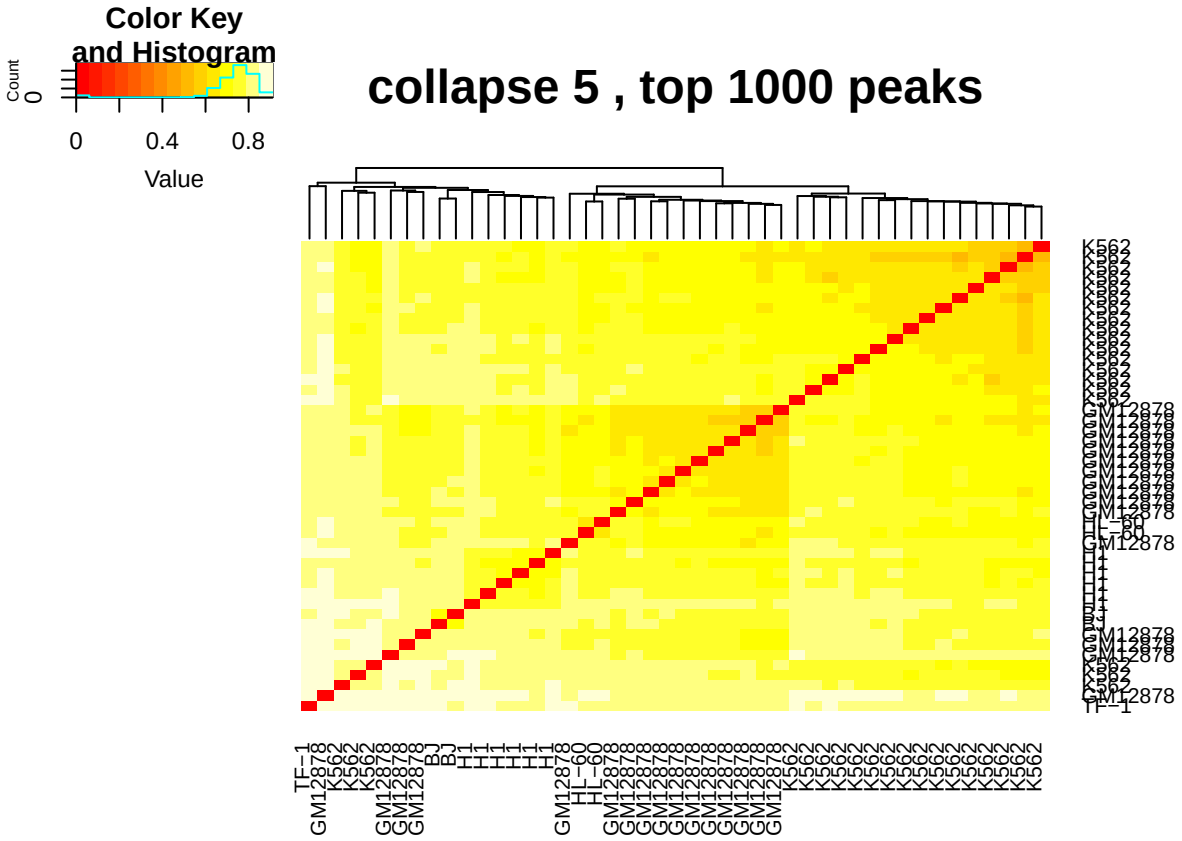
```
coll <- 5
fn <- paste("/Users/zhixianglin/Documents/collaboration/mahdi/scATAC/CellLines6/dataColl",coll, ".rda",
load(fn)
data <- dataColl5$ForGroundC
cells <- dataColl5$SampleSubC
print(table(data[,1]))
```

```
##
##      0      1      2      3      4      5      6      7      8      9
## 130107    535   6662    256   2559    131   1162     78    621     39
##     10     11     12     13     14     15     16     17     18     19
##    333     25    142     10     79      8     46      6     26      2
##     20     21     22     24     26     31     96
##      9      1      5      3      2      1      1
```

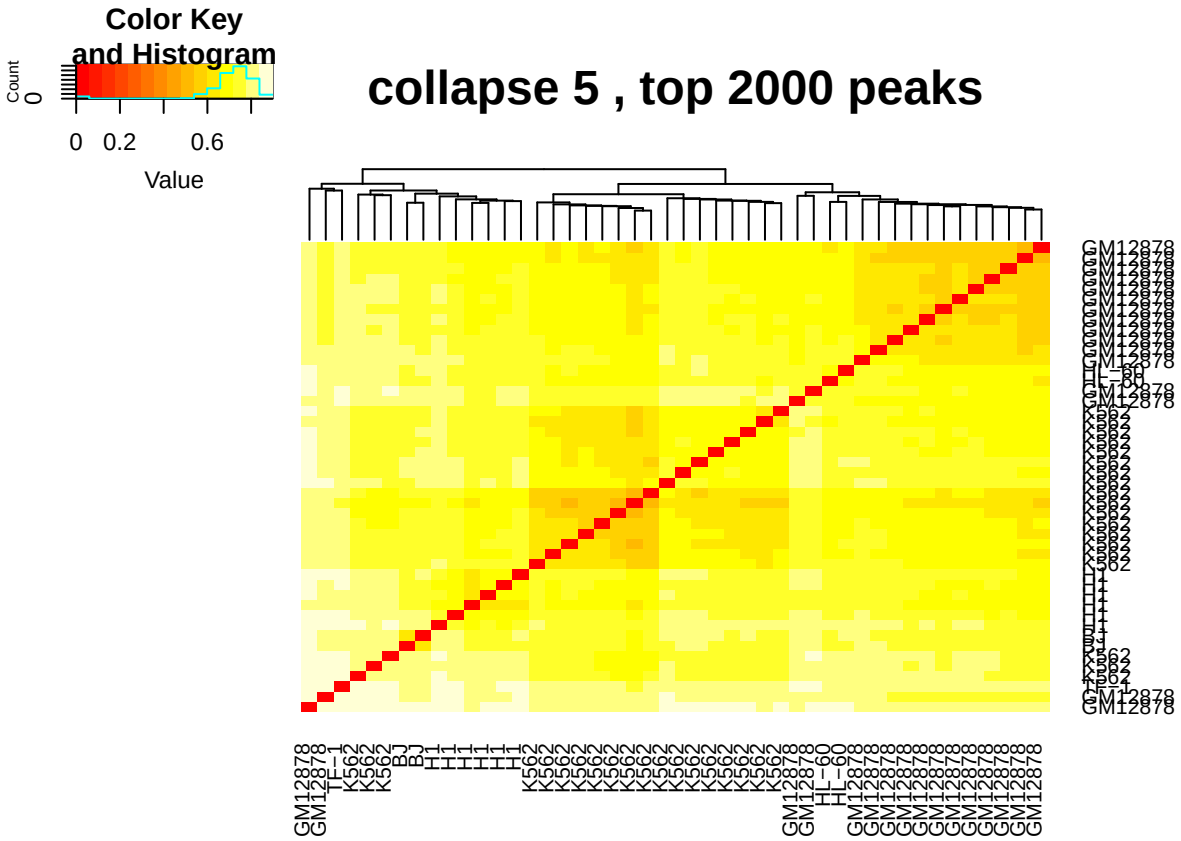
```
sub <- sample(ncol(data), round(ncol(data)/4))
data <- data[,sub]
cells <- cells[sub]
```

```
top <- 1000
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
```

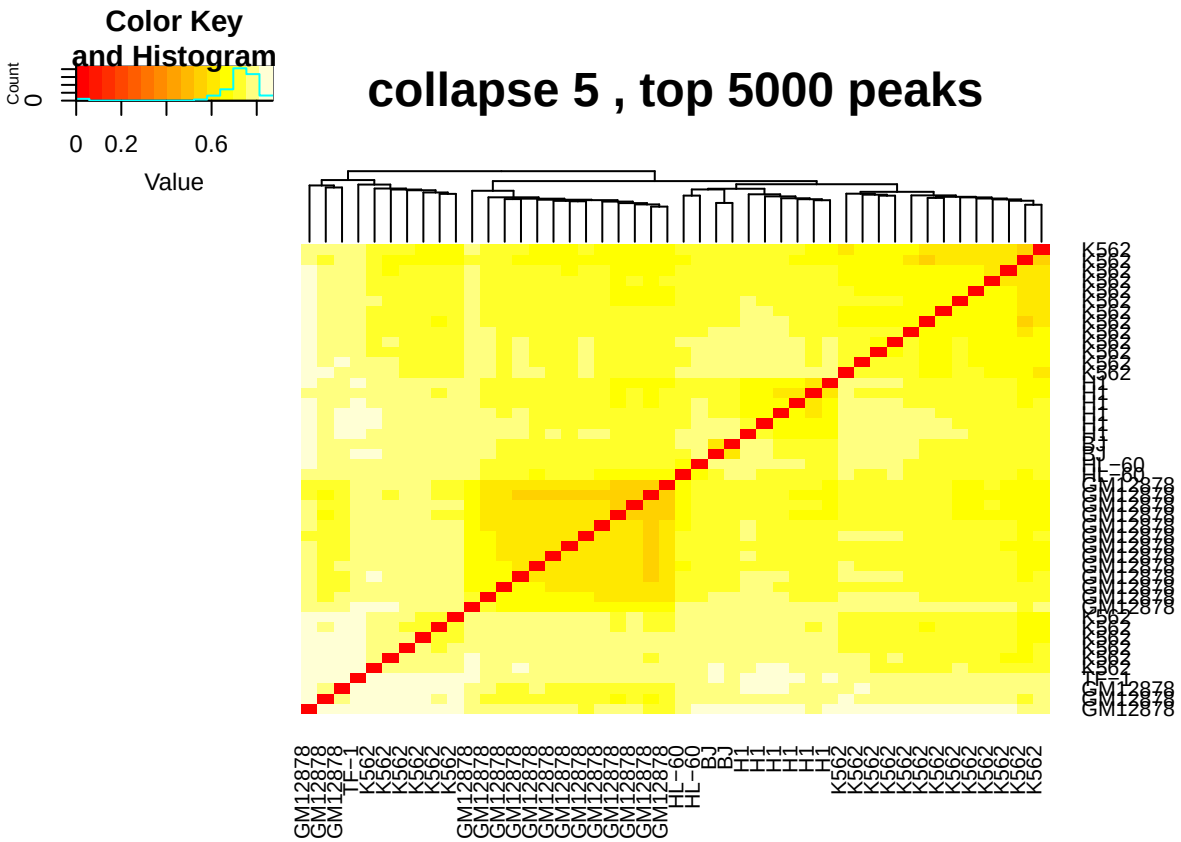
```
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top", t
```




```
top <- 2000
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top", t
```




```
top <- 5000
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top", 
```



Do not collapse, pick a subset of cells

```
set.seed(123)
fn <- paste("/Users/zhixianglin/Documents/collaboration/mahdi/scATAC/CellLines6/dataSub30.rda")
load(fn)
names(dataSub30)
```

```
## [1] "ForGroundSub"  "BackGroundSub" "SampleSub"     "spots"
```

```
data <- dataSub30$ForGroundSub
cells <- dataSub30$SampleSub
```

The number of non-zeros

```
sub <- sample(ncol(data), 50)
data <- data[,sub]
cells <- cells[sub]
colSums(data!=0)
```

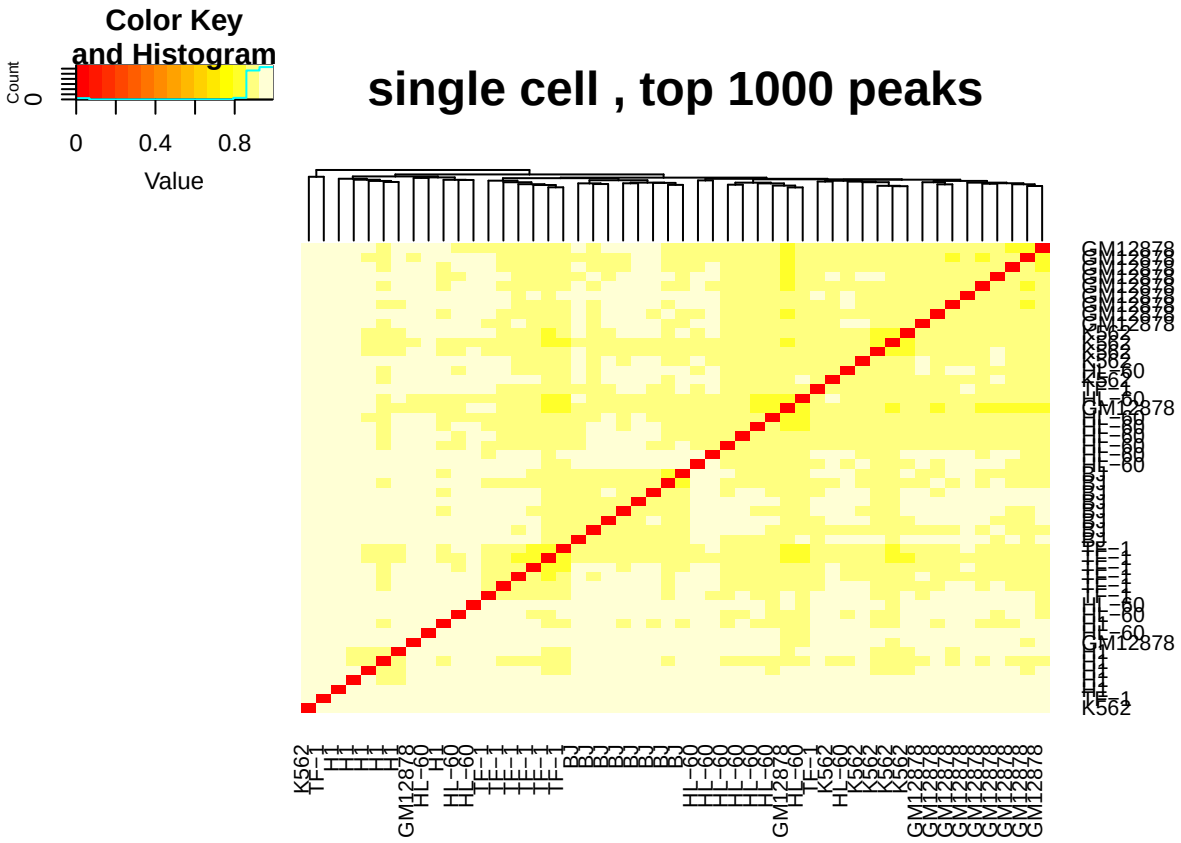
```
##  X724  X665  X955 X1003  X732 X1408 X1049  X154  X447 X1636 X1004  X566
##  5654  3546  1764  2220 19846  2069  4503  1135  4641  2191  5475  2250
## X1172  X243 X1596  X404  X185  X782  X152  X433   X14 X1189  X428 X1643
## 16048  2728  4398  2535  3208  3385  2322  3486  1658  3711  7427  3799
## X1550  X160 X1461  X175  X603  X748 X1374  X899  X671  X619 X1480 X1712
##  5099 13290  2893  2100  1173  2412  5735  3361  3091  3141  3947  1892
## X1042 X1664  X183 X1210 X1184 X1173 X1608   X60  X890  X461 X1582  X221
```

```
##  5480  3010  2471  2301  2689  3411  2035  2419  3797  4803  9587  4328
##  X848   X53
##  2794  3904
```

```
min(colSums(data!=0))
```

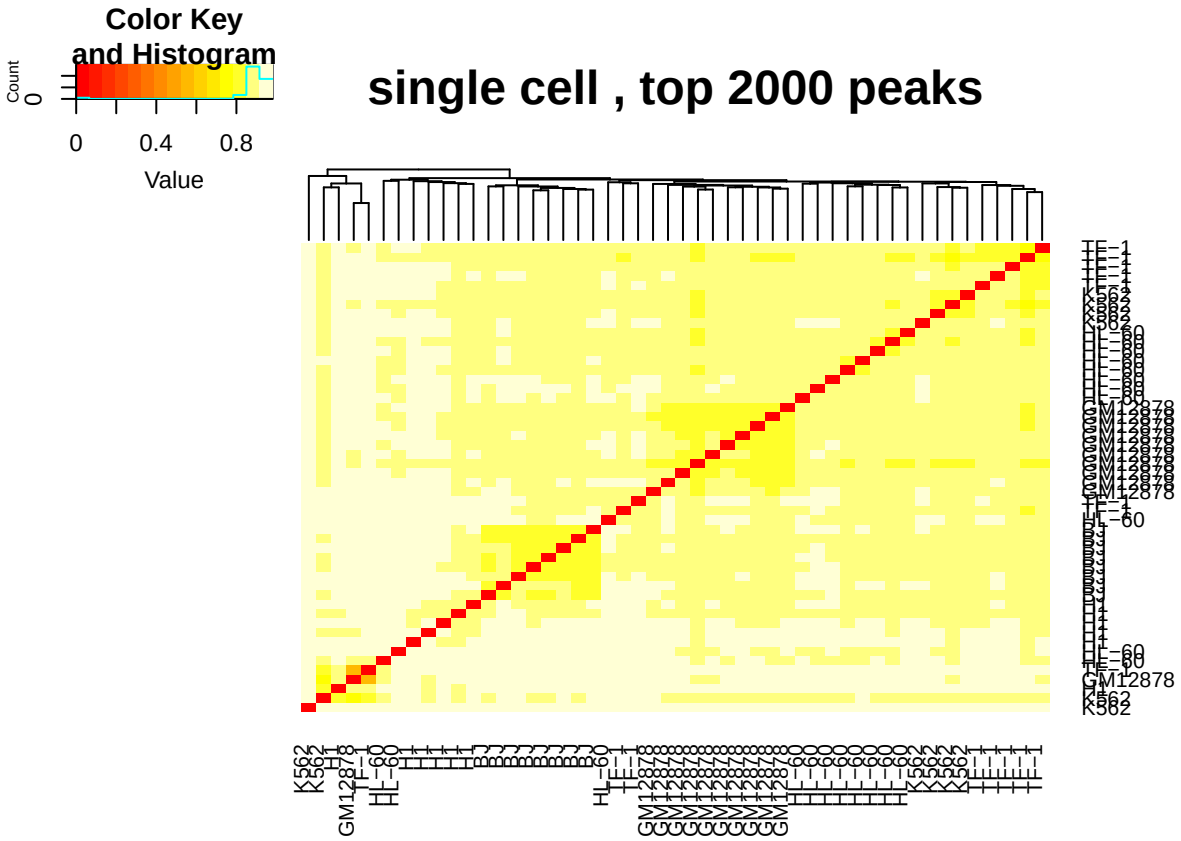
```
## [1] 1135
```

```
top <- 1000
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("single cell", ", top", top
```

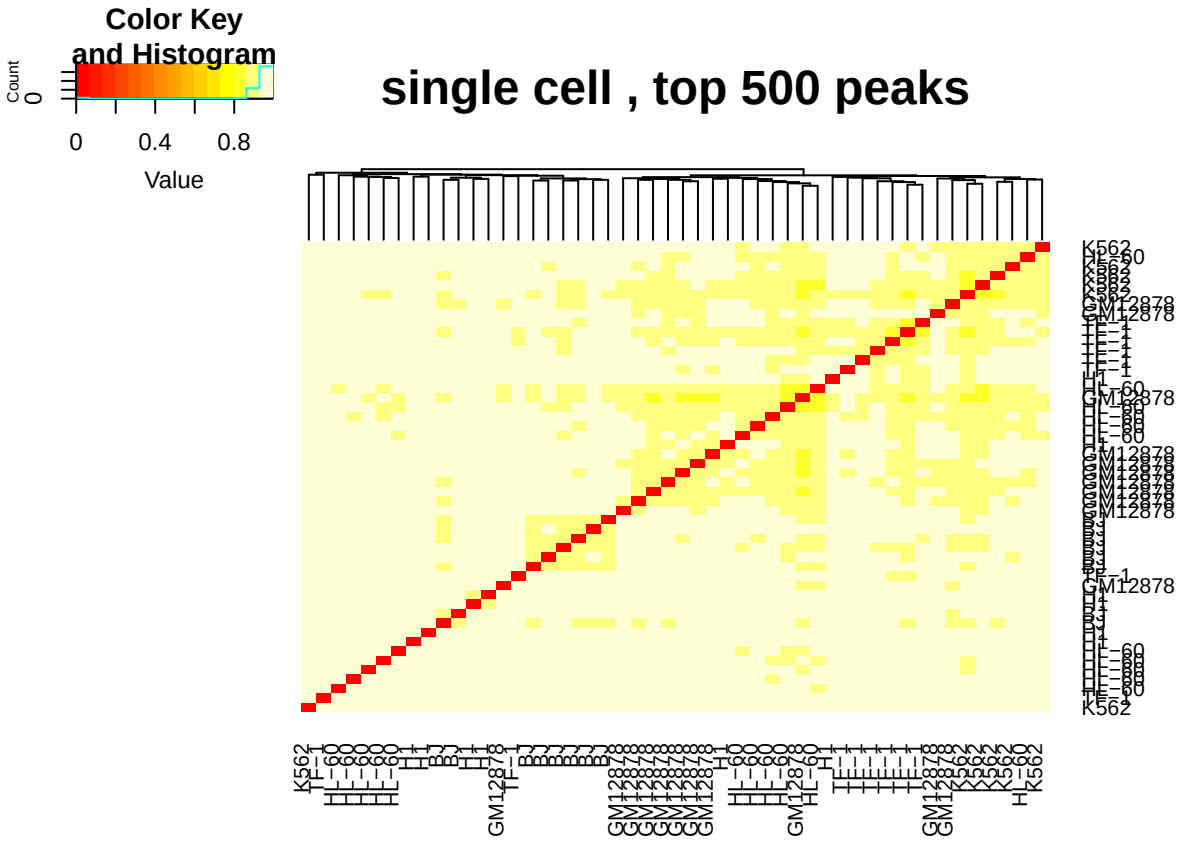



```
top <- 2000
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
```

```
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("single cell", ", top", top
```




```
top <- 500
distM <- matrix(0, ncol(data), ncol(data))
for (i in 1:(ncol(data)-1)){
  for (j in (i+1):ncol(data)){
    distM[i,j] <- 1 - caloverlap(data[,i],data[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("single cell", ", top", top
```



Clearly, when more cells are collapsed, the clusters become more separated.

Idea: modification of the k-means algorithm, instead of calculating the mean, use some robust statistics for calculating the distance. Basicly we need to aggregate cells to lower the noise level, while allowing dissimilar cells to go out/similar cells to go in. Also applies to Drop-Seq data.

Use fold-change

Calculate window size

```
spots <- dataSub30$spots
ws <- spots[,3] - spots[,2] + 1
```

Calculate fold-change, collapse 20

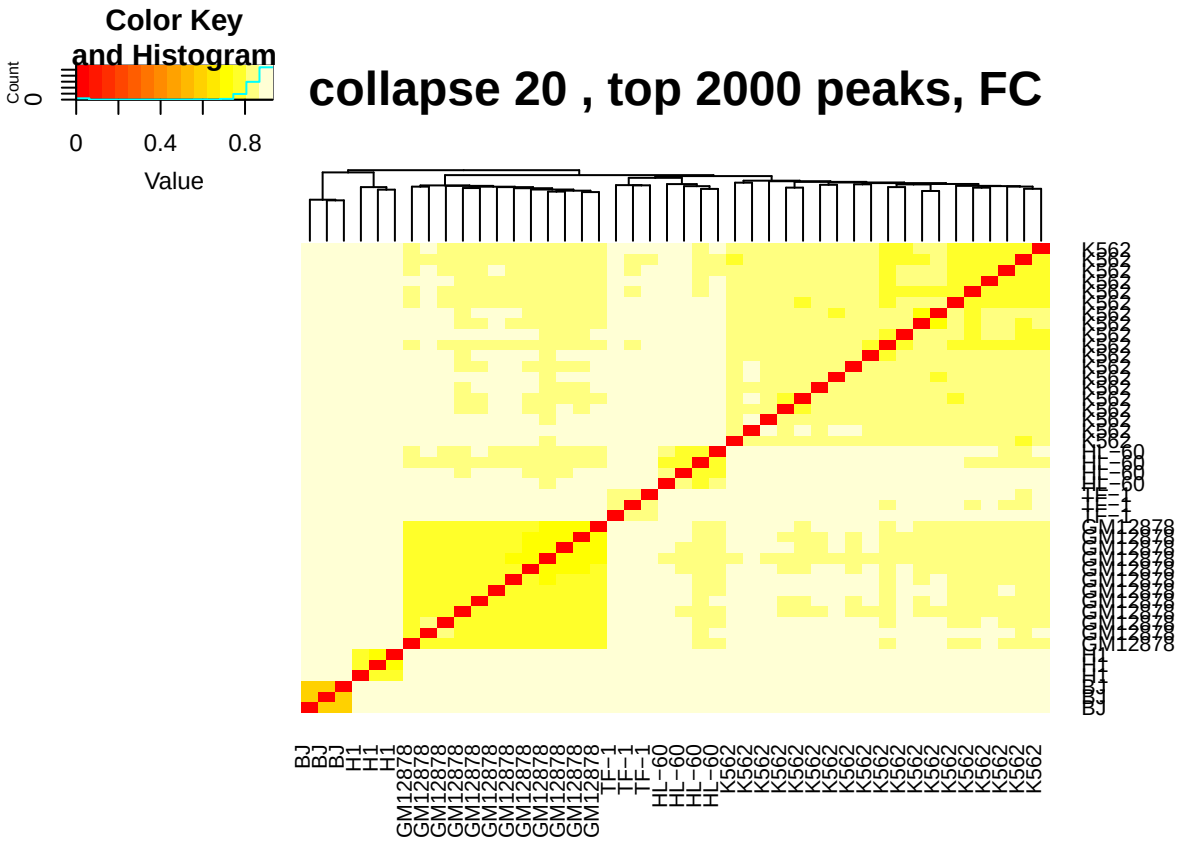
```
coll <- 20
top <- 2000
fn <- paste("/Users/zhixianglin/Documents/collaboration/mahdi/scATAC/CellLines6/dataColl",coll, ".rda",
load(fn)
dataF <- dataColl20$ForGroundC
dataB <- dataColl20$BackGroundC
cells <- dataColl20$SampleSubC
## filter low background
Bmin <- apply(dataB, 1, min)
```

```
thres <- 10
peak_sub <- which(Bmin>=thres)
print(length(peak_sub))
```

```
## [1] 140545
```

```
dataF <- dataF[peak_sub,]
dataB <- dataB[peak_sub,]
## scale to make sure the fold change not too large/small
fc <- dataF/dataB/ws[peak_sub]
fc <- fc/median(fc[fc!=0])

distM <- matrix(0, ncol(fc), ncol(fc))
for (i in 1:(ncol(fc)-1)){
  for (j in (i+1):ncol(fc)){
    distM[i,j] <- 1 - caloverlap(fc[,i],fc[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top", 
```



Calculate fold-change, collapse 10

```
coll <- 10
top <- 2000
```

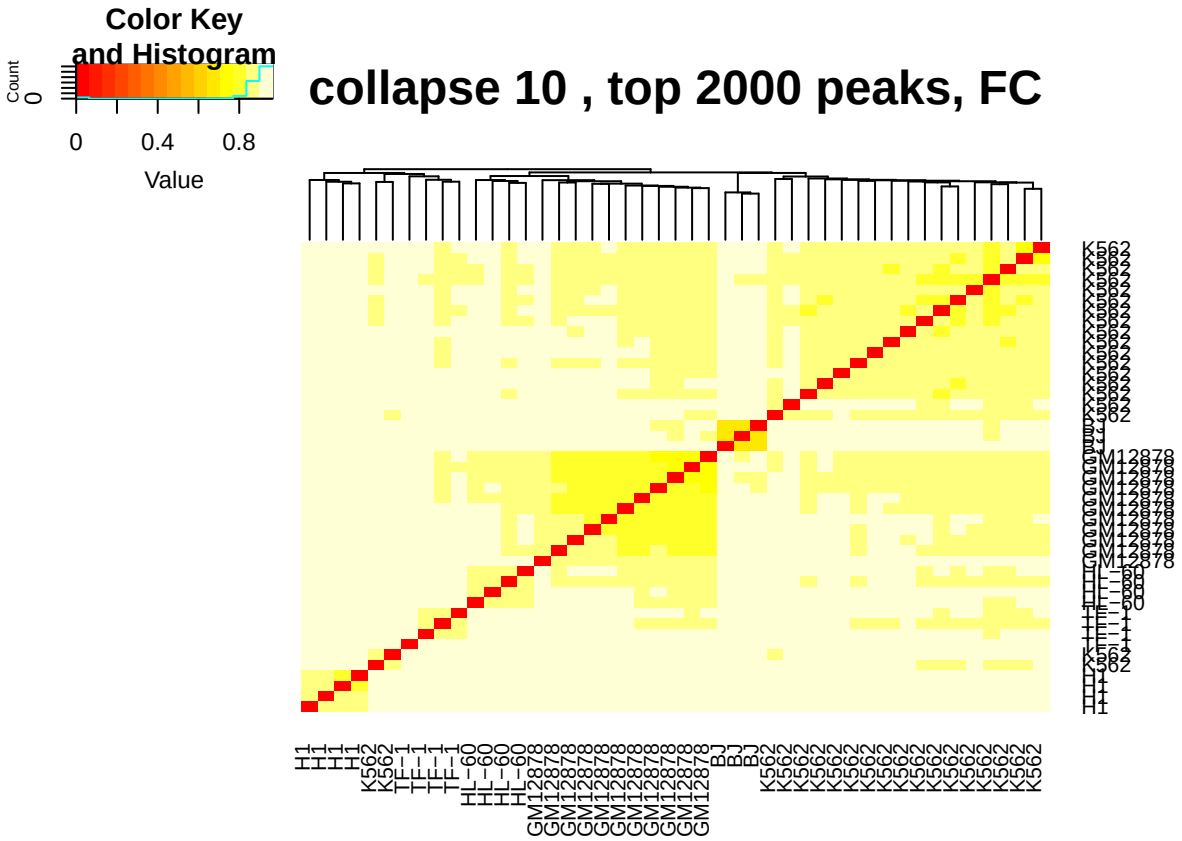
```
fn <- paste("/Users/zhixianglin/Documents/collaboration/mahdi/scATAC/CellLines6/dataColl",coll, ".rda",
load(fn)
dataF <- dataColl10$ForGroundC
dataB <- dataColl10$BackGroundC
cells <- dataColl10$SampleSubC
## filter low background
Bmin <- apply(dataB, 1, min)
thres <- 10
peak_sub <- which(Bmin>=thres)
dataF <- dataF[peak_sub,]
dataB <- dataB[peak_sub,]
print(length(peak_sub))
```

```
## [1] 130650
```

```
sub <- sample(ncol(dataF), round(ncol(dataF)/2))
dataF <- dataF[,sub]
dataB <- dataB[,sub]
cells <- cells[sub]

## scale to make sure the fold change not too large/small
fc <- dataF/dataB/ws[peak_sub]
fc <- fc/median(fc[fc!=0])

distM <- matrix(0, ncol(fc), ncol(fc))
for (i in 1:(ncol(fc)-1)){
  for (j in (i+1):ncol(fc)){
    distM[i,j] <- 1 - caloverlap(fc[,i],fc[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top",
```



Calculate fold-change, collapse 5

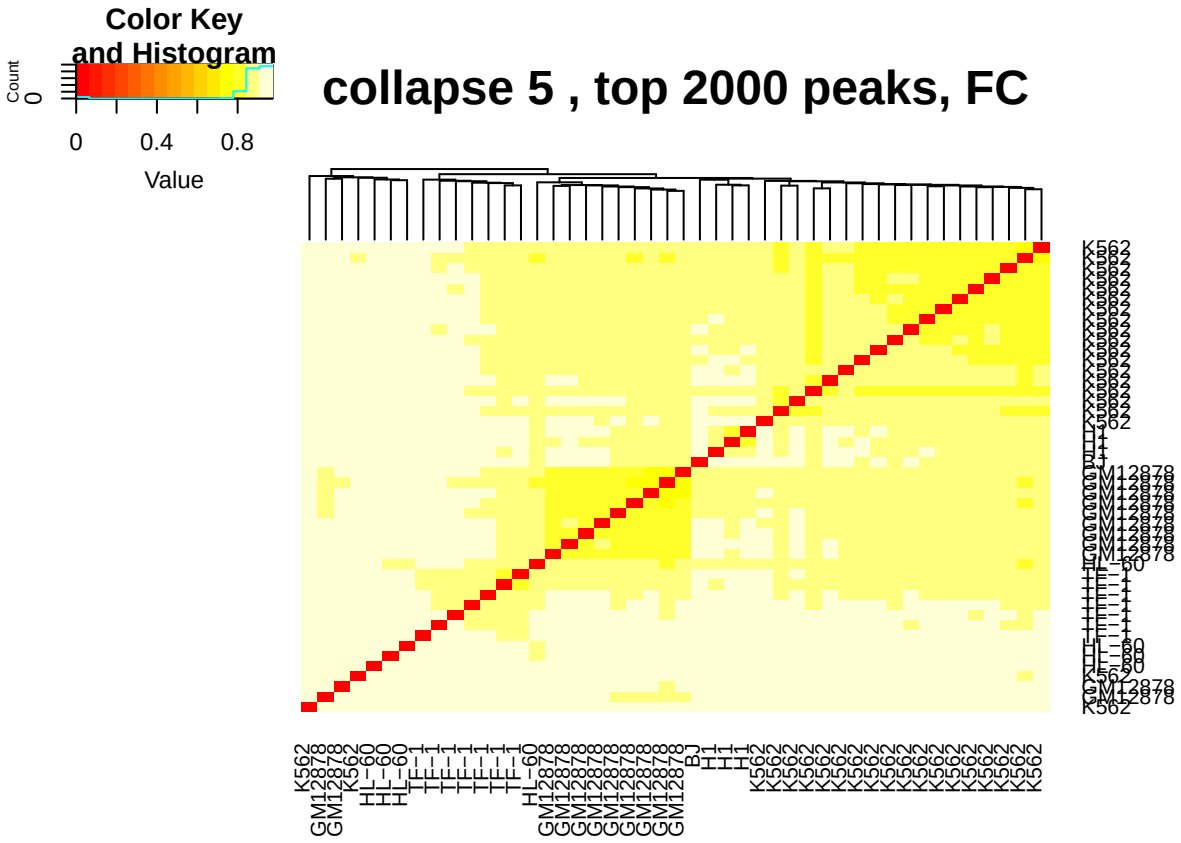
```
coll <- 5
top <- 2000
fn <- paste("/Users/zhixianglin/Documents/collaboration/mahdi/scATAC/CellLines6/dataColl",coll, ".rda",
load(fn)
dataF <- dataColl5$ForGroundC
dataB <- dataColl5$BackGroundC
cells <- dataColl5$SampleSubC
## filter low background
Bmin <- apply(dataB, 1, min)
thres <- 10
peak_sub <- which(Bmin>=thres)
dataF <- dataF[peak_sub,]
dataB <- dataB[peak_sub,]
print(length(peak_sub))
```

```
## [1] 100410
```

```
sub <- sample(ncol(dataF), round(ncol(dataF)/4))
dataF <- dataF[,sub]
dataB <- dataB[,sub]
cells <- cells[sub]

## scale to make sure the fold change not too large/small
fc <- dataF/dataB/ws[peak_sub]
fc <- fc/median(fc[fc!=0])
```

```
distM <- matrix(0, ncol(fc), ncol(fc))
for (i in 1:(ncol(fc)-1)){
  for (j in (i+1):ncol(fc)){
    distM[i,j] <- 1 - caloverlap(fc[,i],fc[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("collapse", coll, ", top", t
```



Calculate fold-change, single cell, unreliable background

```
dataF <- dataSub30$ForGroundSub
dataB <- dataSub30$BackGroundSub
cells <- dataSub30$SampleSub
```

```
sub <- sample(ncol(data), 50)
dataF <- dataF[,sub]
dataB <- dataB[,sub]
cells <- cells[sub]
## filter low background
Bmin <- apply(dataB, 1, min)
thres <- 5
peak_sub <- which(Bmin>=thres)
dataF <- dataF[peak_sub,]
```

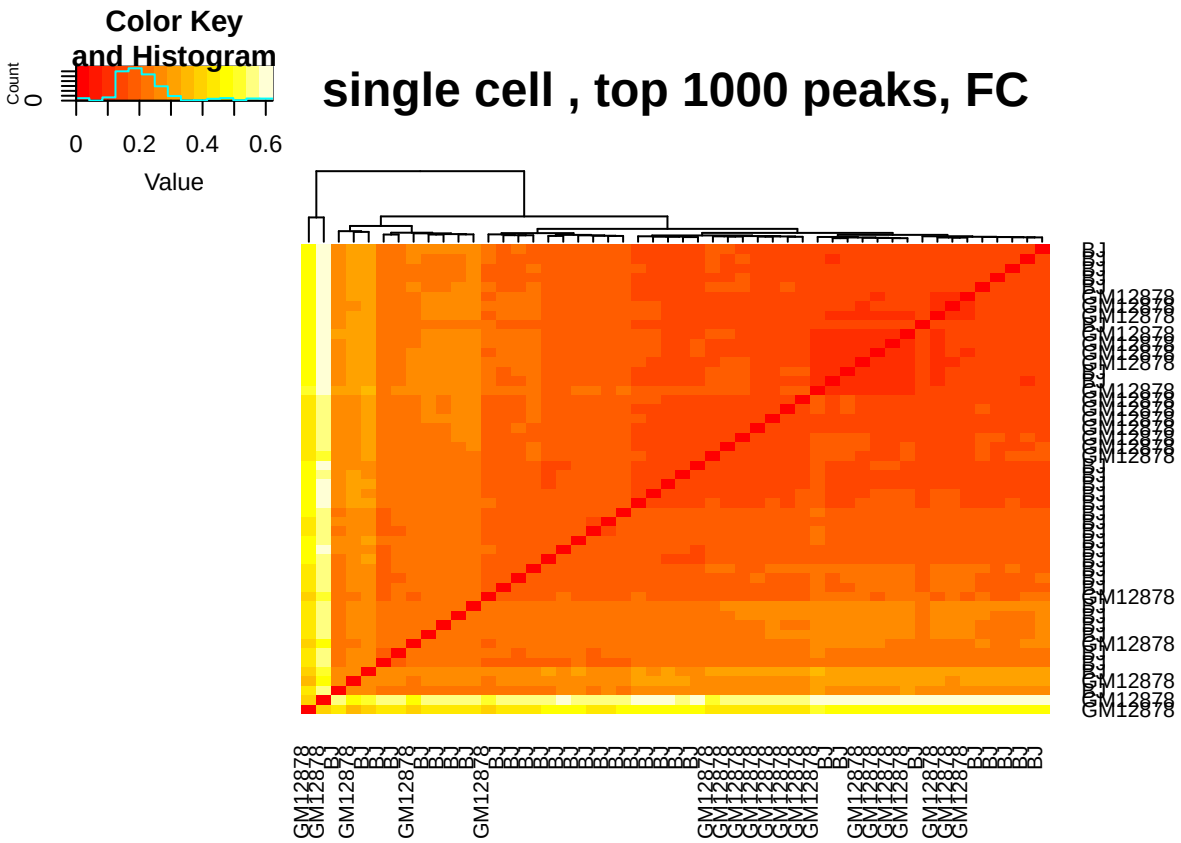
```
dataB <- dataB[peak_sub,]
print(length(peak_sub))
```

```
## [1] 10580
```

```
colSums(dataF!=0)
```

```
##  X890  X485 X1589  X341 X1431  X183 X1480  X192 X1664   X36 X1184 X1185
##   559   496   535   288   671   347   585   725   437   342   421   389
## X1699 X1586  X155 X1383 X1408  X848  X782  X517  X699  X461 X1700  X305
##   511   499   859  1574   291   438   551   299   805   696   369   398
##  X776  X266   X33  X581 X1129  X756  X603  X308  X185  X707  X748 X1542
##   457   360   358   620   723   437   162   540   428   283   355   376
## X1382  X740 X1582  X704  X335  X835  X722 X1326 X1596 X1337 X1210  X702
##   573   669  1268   443   463   423   578   405   679   687   330   437
##  X780  X409
##   333   251
```

```
fc <- dataF/dataB/ws[peak_sub]
fc <- fc/median(fc[fc!=0])
distM <- matrix(0, ncol(fc), ncol(fc))
for (i in 1:(ncol(fc)-1)){
  for (j in (i+1):ncol(fc)){
    distM[i,j] <- 1 - caloverlap(fc[,i],fc[,j],top)
  }
}
top <- 1000
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("single cell", ", top", top
```



Use peak length for the adjustment

Collapse 20 cells

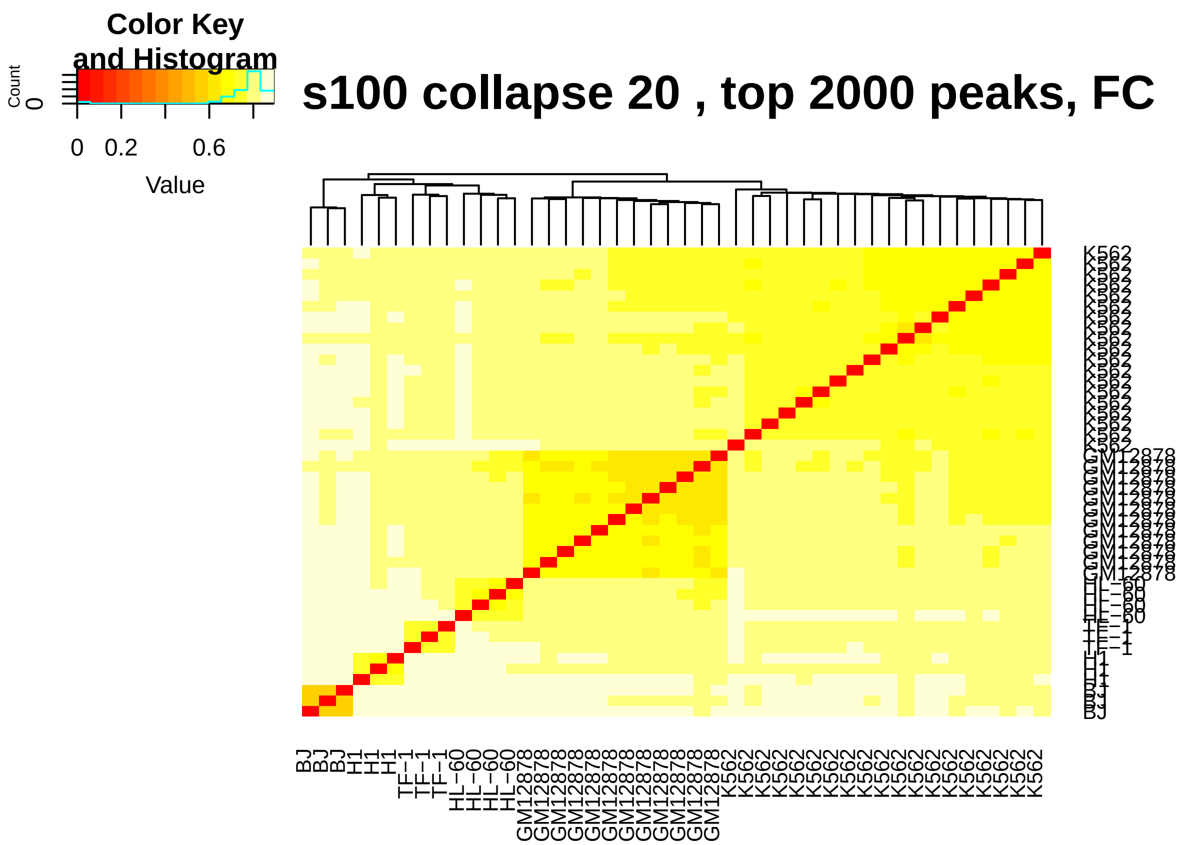
```
coll <- 20
top <- 2000
fn <- paste("/Users/zhixianglin/Documents/collaboration/mahdi/scATAC/CellLines6/dataColl",coll, ".rda",
load(fn)
dataF <- dataColl20$ForGroundC
dataB <- dataColl20$BackGroundC
cells <- dataColl20$SampleSubC
## filter low background
Bmin <- apply(dataB, 1, min)
thres <- 100
peak_sub <- which(Bmin>=thres)
print(length(peak_sub))
```

```
## [1] 101700
```

Fold-change, more stringent background threshold

```
dataF <- dataF[peak_sub,]
dataB <- dataB[peak_sub,]
## scale to make sure the fold change not too large/small
fc <- dataF/dataB/ws[peak_sub]
fc <- fc/median(fc[fc!=0])
```

```r
distM <- matrix(0, ncol(fc), ncol(fc))
for (i in 1:(ncol(fc)-1)){
  for (j in (i+1):ncol(fc)){
    distM[i,j] <- 1 - caloverlap(fc[,i],fc[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("Thres100 collapse", coll,
```

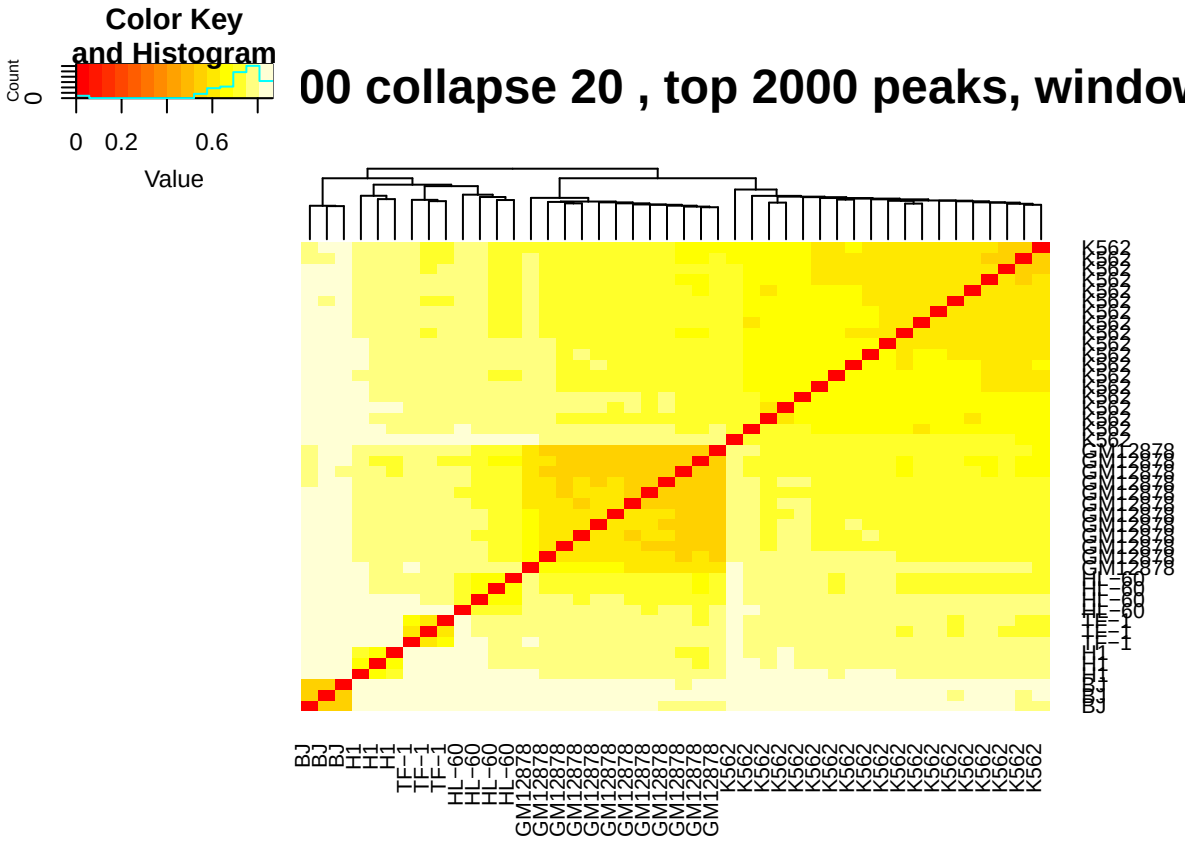


Peak length adjustment, more stringent background threshold

```r
## scale to make sure the fold change not too large/small
fc <- dataF/ws[peak_sub]
fc <- fc/median(fc[fc!=0])

distM <- matrix(0, ncol(fc), ncol(fc))
for (i in 1:(ncol(fc)-1)){
  for (j in (i+1):ncol(fc)){
    distM[i,j] <- 1 - caloverlap(fc[,i],fc[,j],top)
  }
}
distM <- distM + t(distM)
colnames(distM) <- cells
rownames(distM) <- cells
```

```
heatmap.2(distM, symm = TRUE, dendrogram="column", trace="none", main=paste("Thres100 collapse", coll,
```



We need a good background estimate