automatically removing high missing patients or features

wei zou
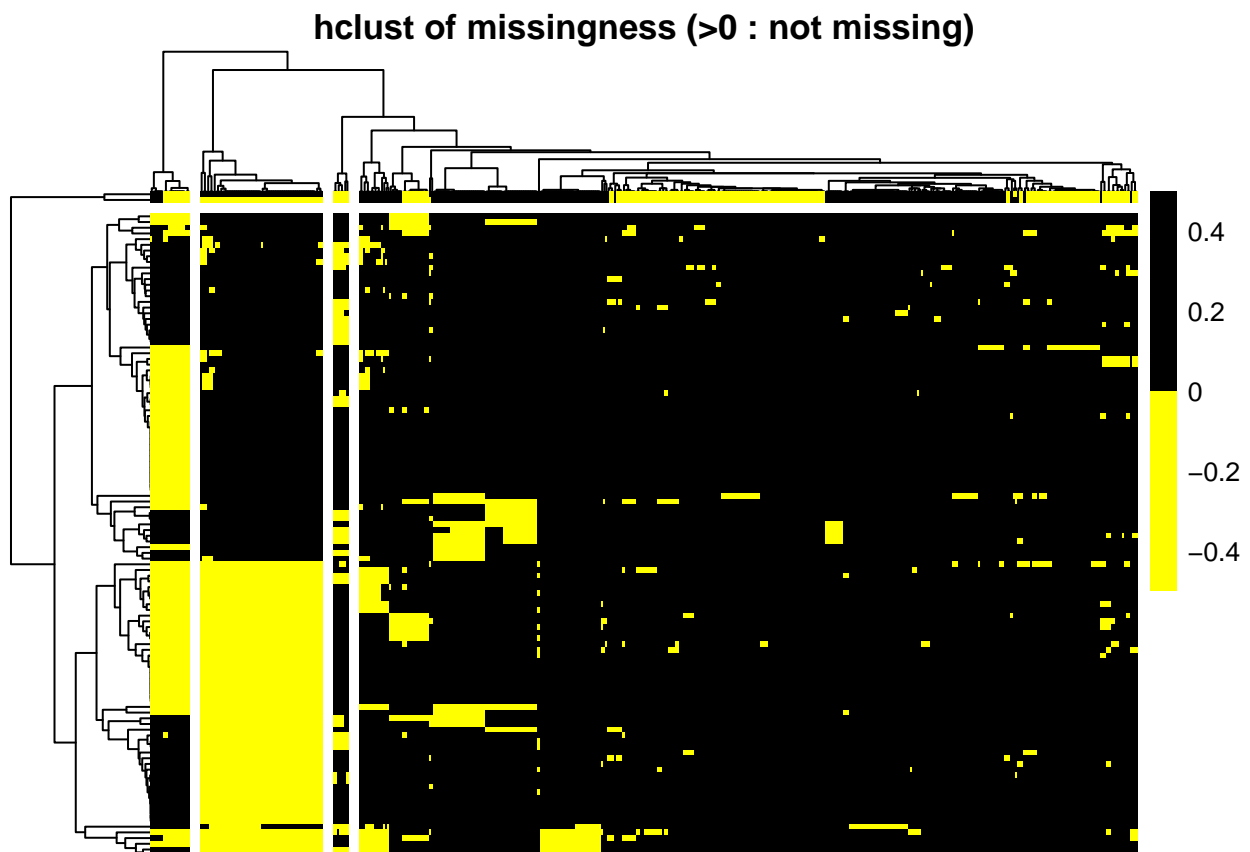
2022-02-10 09:55:01

using Jaccard distance

```r
c1 <- clusterMiss(ds1, dis_measure = list('row'='jaccard', 'col'='jaccard'), numCluster = c('row'=2, 'co
```

```
## raw input dimension: 126 520
## dimension after removing full missing cols/rows: 126 484
## matrix used for clustering: 115 438
## using jaccard,jaccard distance
```

```
## Metric: 'jaccard'; comparing: 115 vectors.
```

```
## Metric: 'jaccard'; comparing: 438 vectors.
```


hclust of missingness (>0 : not missing)

```
## [1] "row groups:"
## pt_groups
```
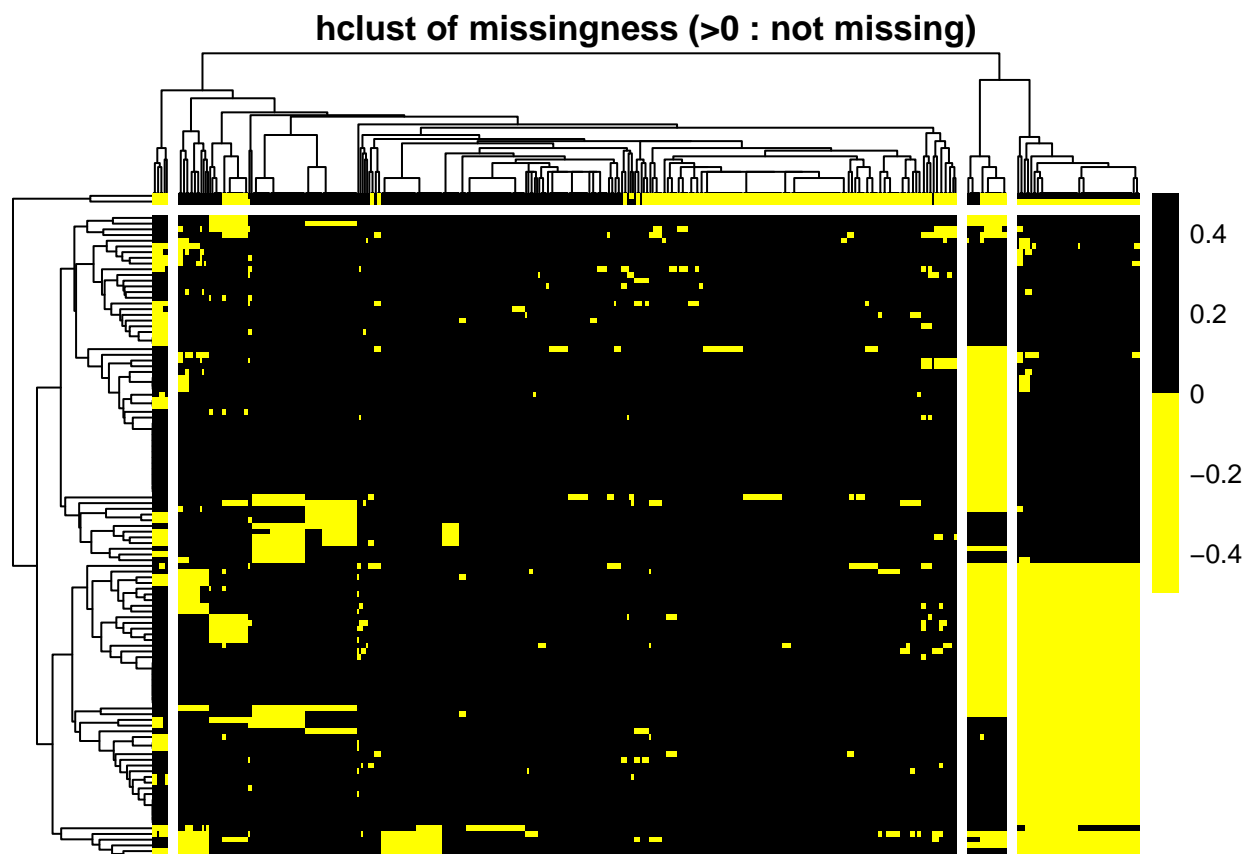
```
##   1   2
## 113   2
## [1] "column groups:"
## ft_groups
##   1   2   3   4
## 357   7  56  18
## [1] "summary of percent non-missing values among excluded items in the original matrix"
## GO28915-264866-401004 GO28915-266618-423025
##             0.6198347             0.5061983
## summary for 81 items
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4683  0.5159  0.5794  0.5684  0.5873  0.7619
## [1] "summary of percent non-missing values after removing patients and samples with high missing"
## summary for 124 items
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8065  0.9404  0.9851  0.9648  1.0000  1.0000
## summary for 403 items
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7823  0.9516  0.9839  0.9648  0.9919  1.0000
```

using eucledian distance

this is an option; but not the default setting

```r
c1 <- clusterMiss(ds1, dis_measure = list('row'='euclidean', 'col'='euclidean'), numCluster = c('row'=2
```

```
## raw input dimension: 126 520
## dimension after removing full missing cols/rows: 126 484
## matrix used for clustering: 115 438
## using euclidean,euclidean distance
```

hclust of missingness (>0 : not missing)


```
## [1] "row groups:"
## pt_groups
##   1   2
## 113   2
## [1] "column groups:"
## ft_groups
##   1   2   3   4
## 357   7  56  18
## [1] "summary of percent non-missing values among excluded items in the original matrix"
## GO28915-264866-401004 GO28915-266618-423025
##             0.6198347             0.5061983
## summary for 81 items
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4683  0.5159  0.5794  0.5684  0.5873  0.7619
## [1] "summary of percent non-missing values after removing patients and samples with high missing"
## summary for 124 items
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8065  0.9404  0.9851  0.9648  1.0000  1.0000
## summary for 403 items
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7823  0.9516  0.9839  0.9648  0.9919  1.0000
```