More on real data analysis

2019/01

Roadmap

We were considering about analysing real microbiome data with the new desinged ADMM and AMA method. However, the previous results were not good enough since both Eric's method and ADMM/AMA did not classfiy the microbiome clearly (the method to generate groups in microbiome types was not efficient). In this document, I am trying to cluster microbiome into groups in better ways. Row color bars are also added in the heatmaps.

The document following:

- 1. Summary of the simulation method
- 2. Results of the pure simulated data with the same method
- 3. Results of the simulated microbiome data with three clusters
- 4. Results of the simulated microbiome data with two clusters

1. Summary of the simulation method

The basic idea is:

- For the Columns

Sample 30 patients in control group, whose gender = female, week = 3; sample 30 patients in PAT group, whose gender = female, week = 3. Combine these two datasets.

- For the Rows

The microbiomes are divided into n groups. Sample x micorbiomes and assign to group 1, sample another x microbiomes and assign to group 2, and so on…

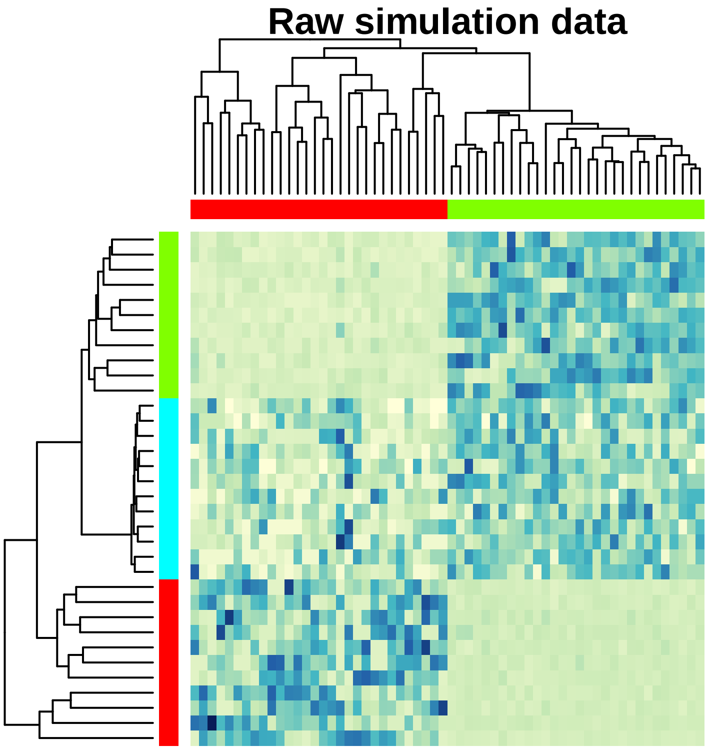
2. Results of the pure simulated data with the same method

To make sure that this kind of method can work, I firstly tried it on a pure simulated data.

Data Generation

```r
# generate column groups, same dimenstion with microbiome data
con = matrix(rnorm(30*34,2,2),34,30)
trl = matrix(rnorm(30*34,4,2),34,30)
dat = cbind(con,trl)

# generate row groups
sam = sample(1:34)
group1 = sam[1:11]
group2 = sam[12:22]
dat[group1,1:30] = dat[group1,1:30] * 8
dat[group2,31:60] = dat[group2,31:60] * 4
```

```
# make it looks more like count data
dat = round(dat)
dat = ifelse(dat<0,0,dat)
for(i in 1:dim(dat)[2]){
  dat[,i] =  dat[,i]/apply(dat,2,sum)[i]
}

# assign group names
colnames(dat) = c(rep('control',30),rep('treatment',30))
rown = rep('group3',34)
rown[group1] = 'group1'
rown[group2] = 'group2'
rownames(dat) = rown

# draw heatmap for the raw data
col_types = colnames(dat)
col_ty = as.numeric(factor(col_types))
row_types = rownames(dat)
row_ty = as.numeric(factor(row_types))
cols <- rainbow(4)
YlGnBu5 <- c('#ffffd9','#c7e9b4','#41b6c4','#225ea8','#081d58')
hmcols <- colorRampPalette(YlGnBu5)(256)

heatmap(dat,col=hmcols,labRow=NA,labCol=NA,
        ColSideCol=cols[col_ty],RowSideCol=cols[row_ty],
        main = 'Raw simulation data')
```



The number of 0s in the dataset:

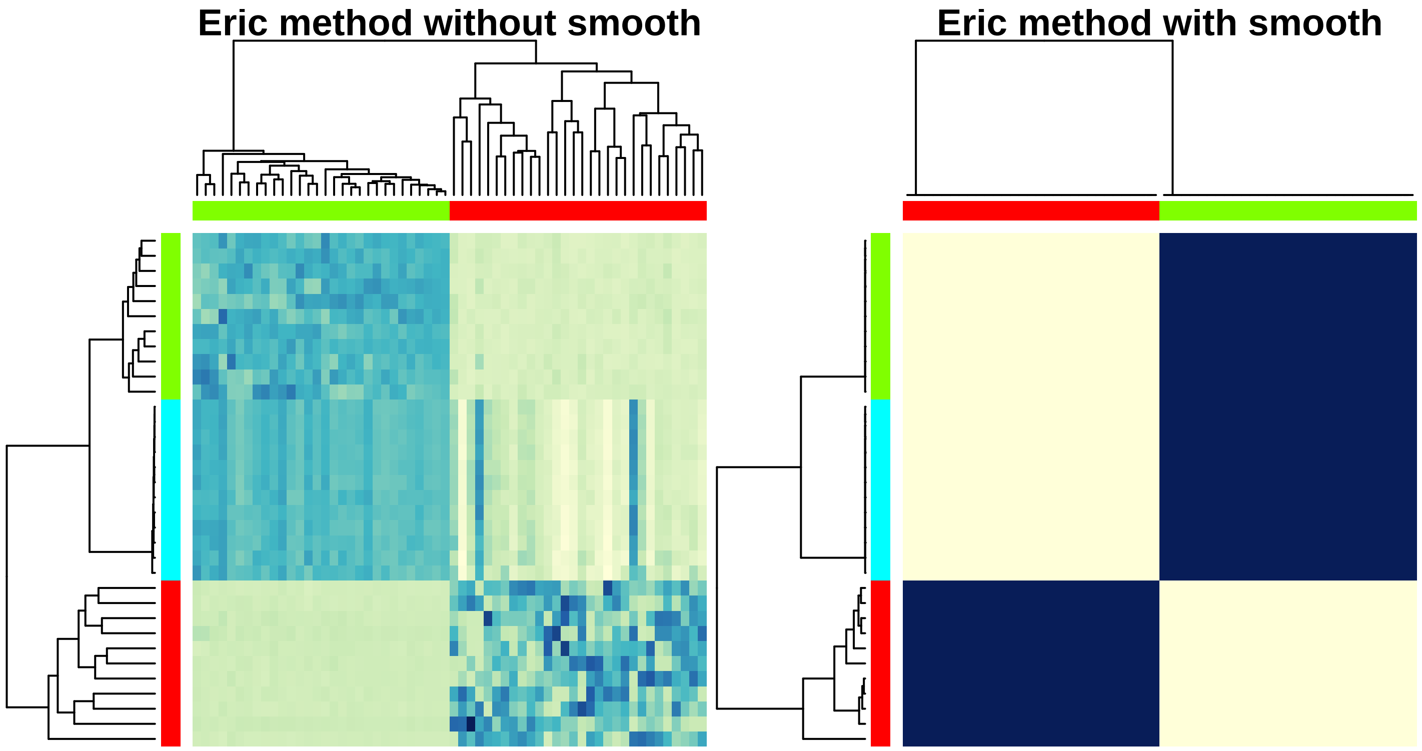
```
sum(abs(dat) < 0.001)
```

```
## [1] 231
```

```
sum(abs(dat) < 0.001)/(34*60)
```

```
## [1] 0.1132353
```

Eric's method

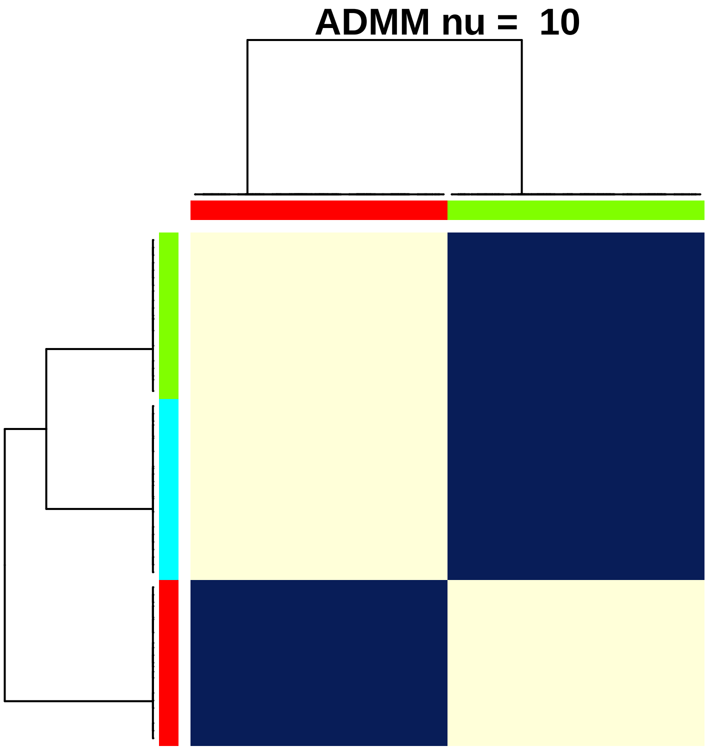


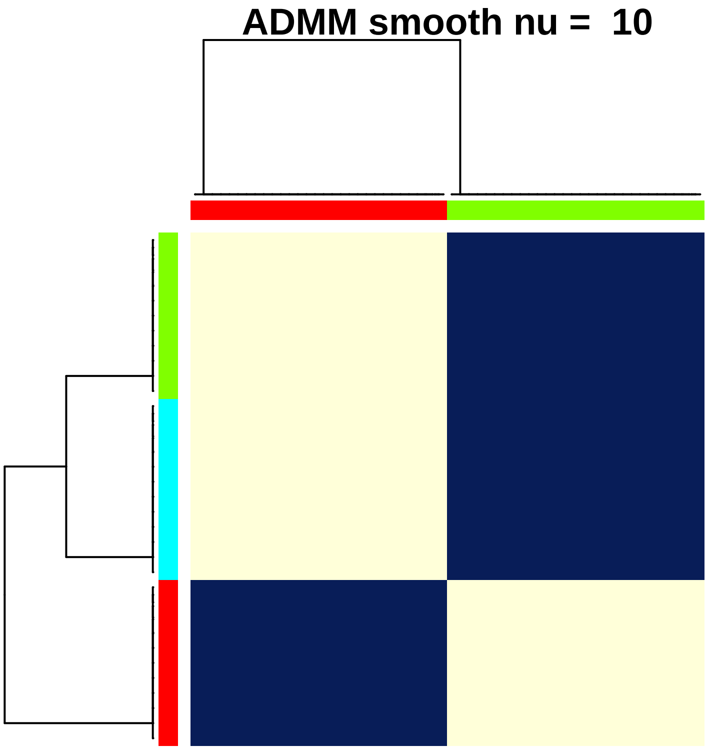
Eric method without smooth

Eric method with smooth

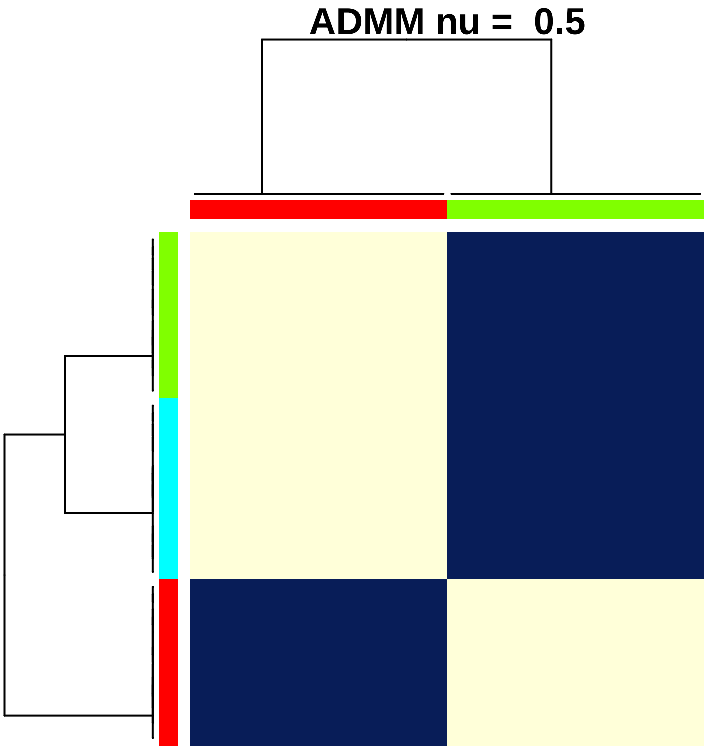
ADMM

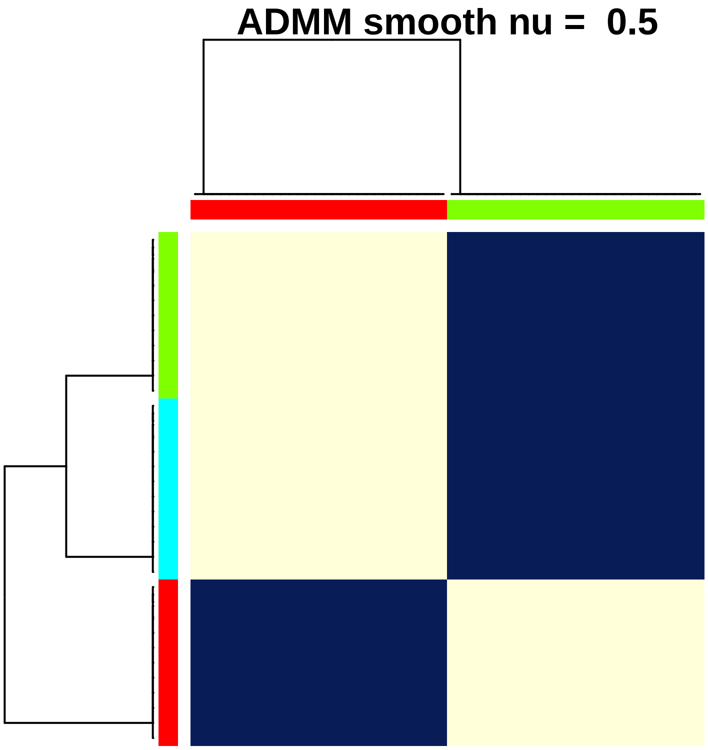
ADMM with the same paramethers (nu1, nu2) as Eric's method:





ADMM with another parameter:





A summary

For the results, we can see that both methods work well. The clusters can be classfied clearly.

ADMM works better than Eric's method. It can work well enought without smooth function.

This kind of method works well for the simulated data. However, maybe not good enought for microbiome data since there are too many 0s.

3. Results of the simulated microbiome data with three clusters

Data preparetion

Columns

Sample 30 patients in control group, whose gender = female, week = 3; sample 30 patients in PAT group, whose gender = female, week = 3. Combine these two datasets.

The dimension of the combined data is:

```r
dim(microbiome)
```

```
## [1] 60 34
```

60 patients and 34 kinds of microbiome.

Rows

The microbiomes are divided into three groups. Sample 11 micorbiomes and assign to group 1, sample another 11 microbiomes and assign to group 2. The rest is group 3.

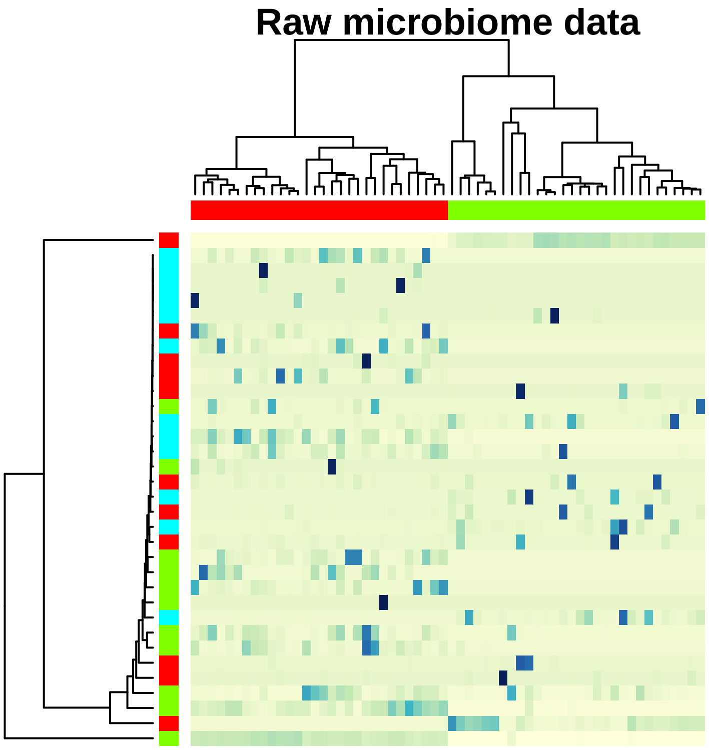
I tried 3 scenarios.

- scenario 1: a. value in [PAT group] and [group 1] times 10. b. value in [control group] and [group 2] times 20
- scenario 2: a. value in [PAT group] and [group 1] times 100. b. value in [control group] and [group 2] times 200
- scenario 3: a. value in [PAT group] and [group 1] times 0.2. b. value in [control group] and [group 2] times 2

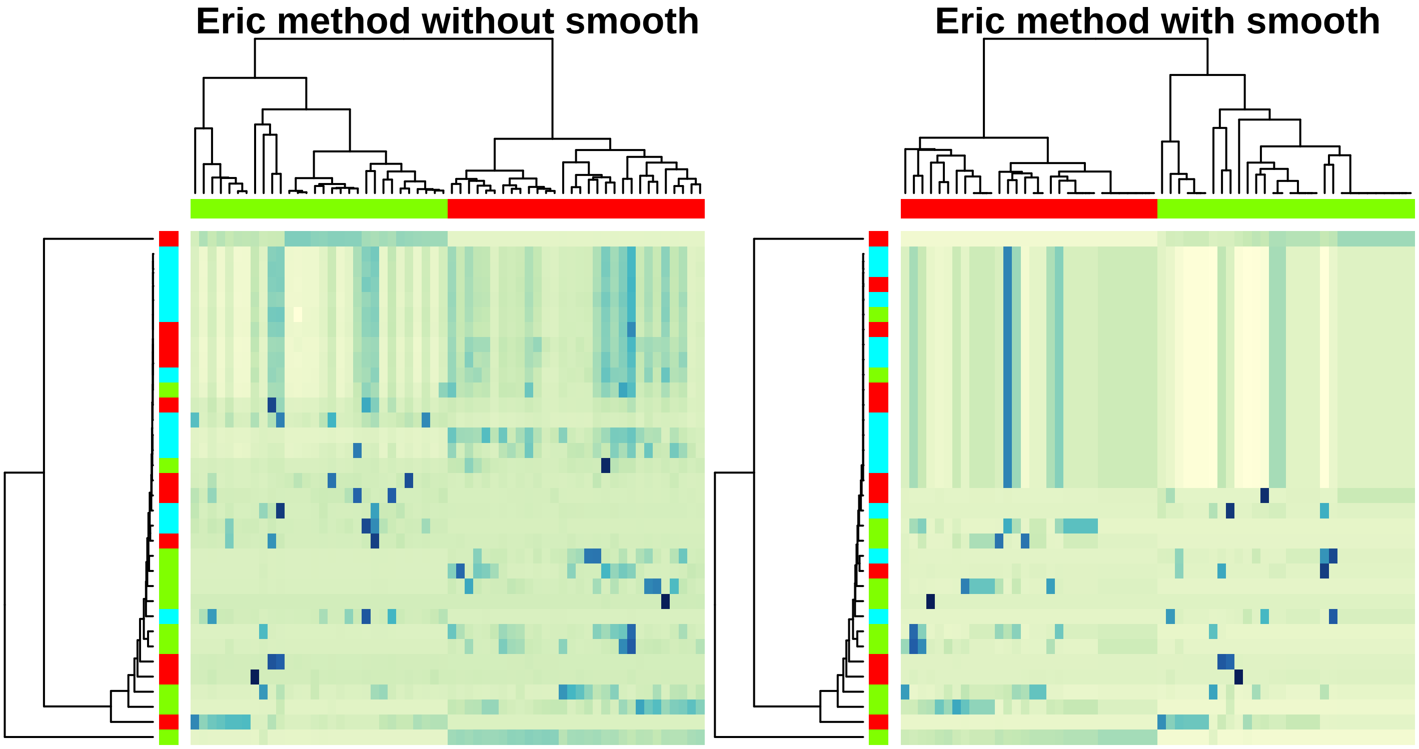
Results

Scenario 1

Heatmap of raw data

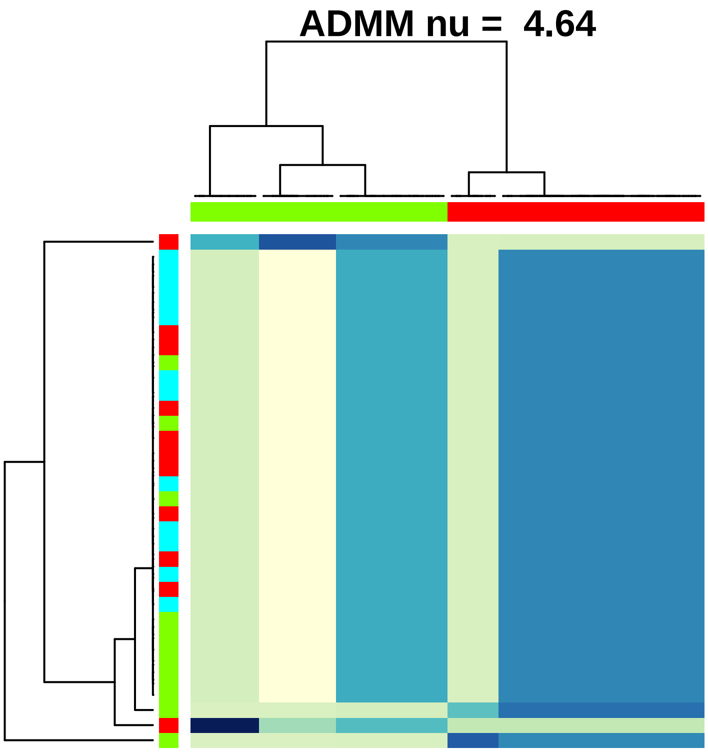


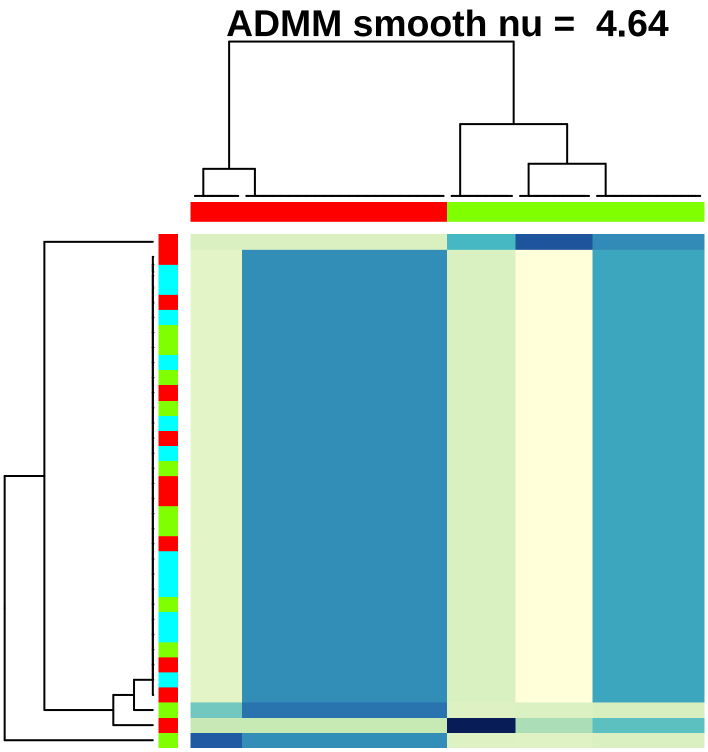
Eric's method



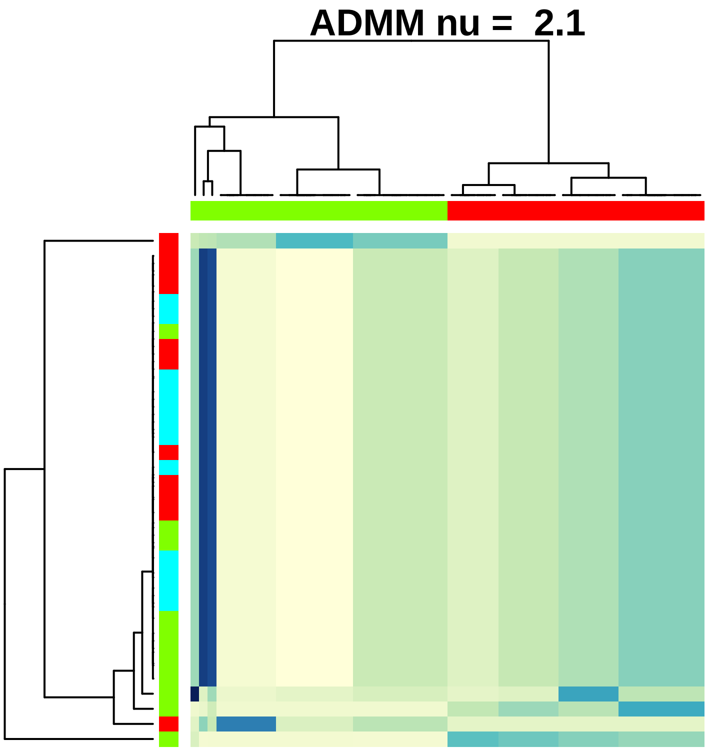
ADMM

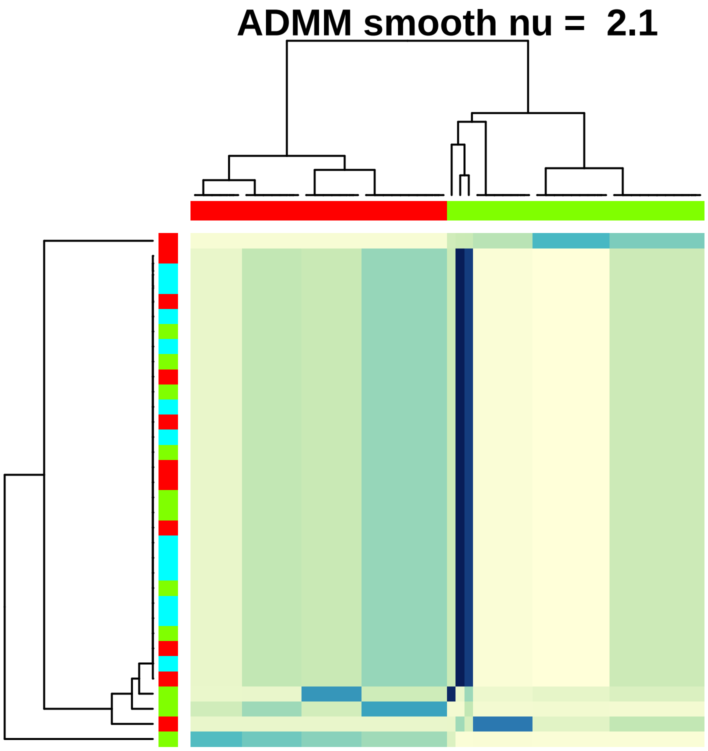
- ADMM with the same parameter (nu1, nu2) with Eric's method





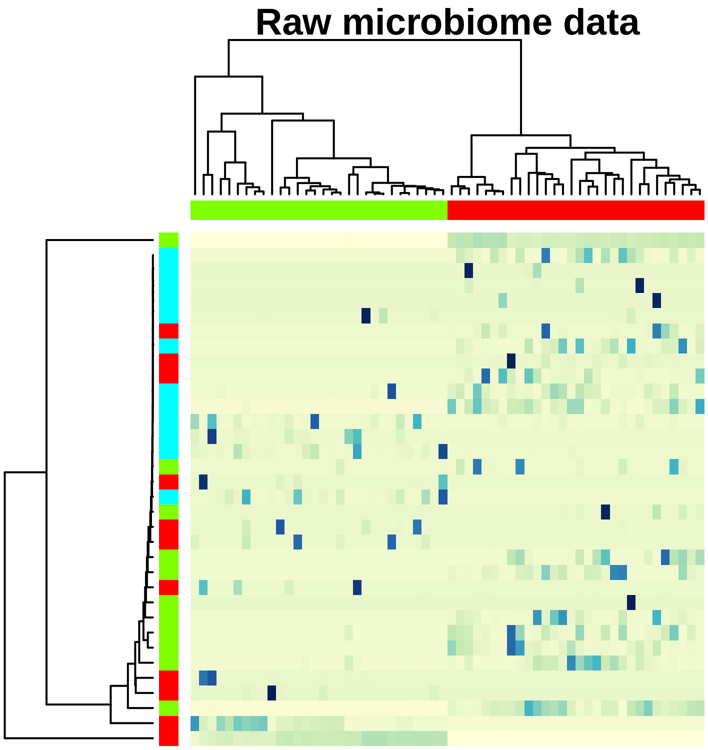
- try another parameter



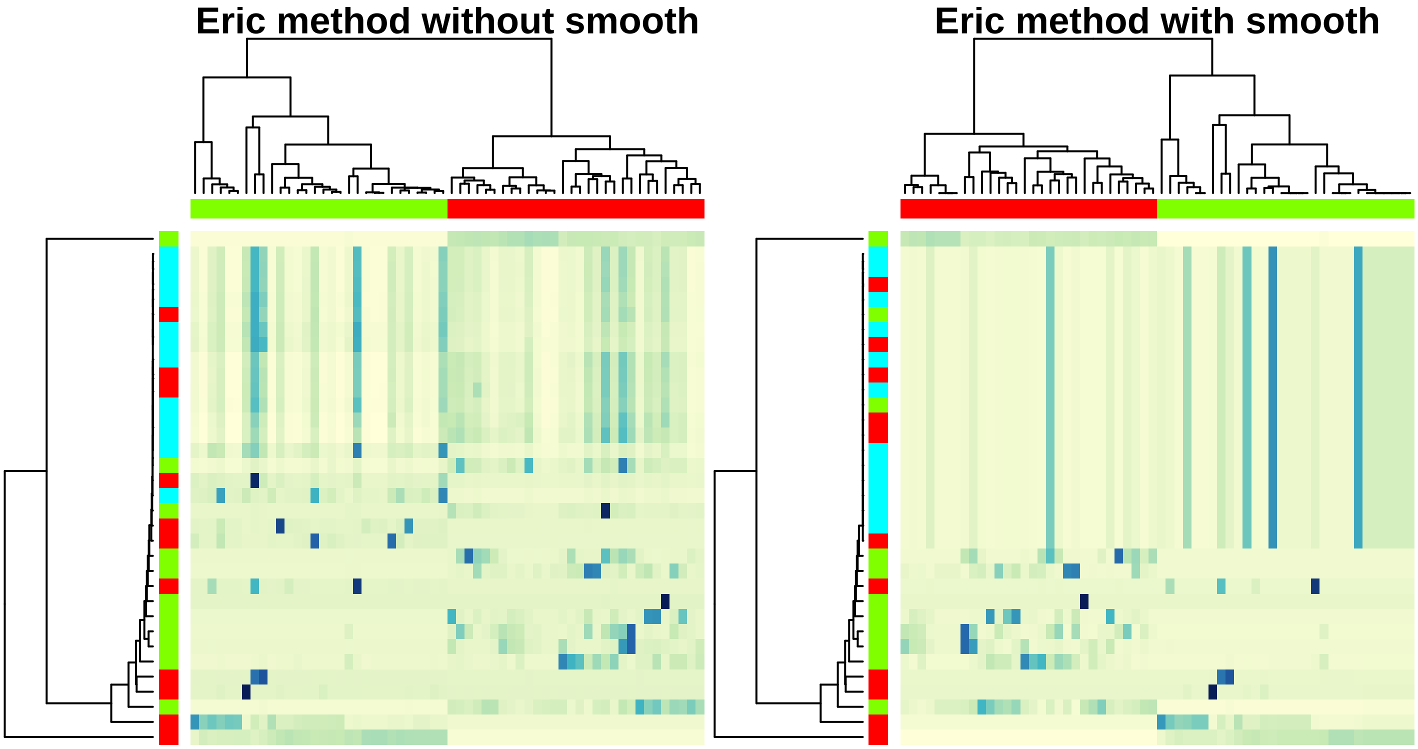


Scenario 2

Heatmap of raw data

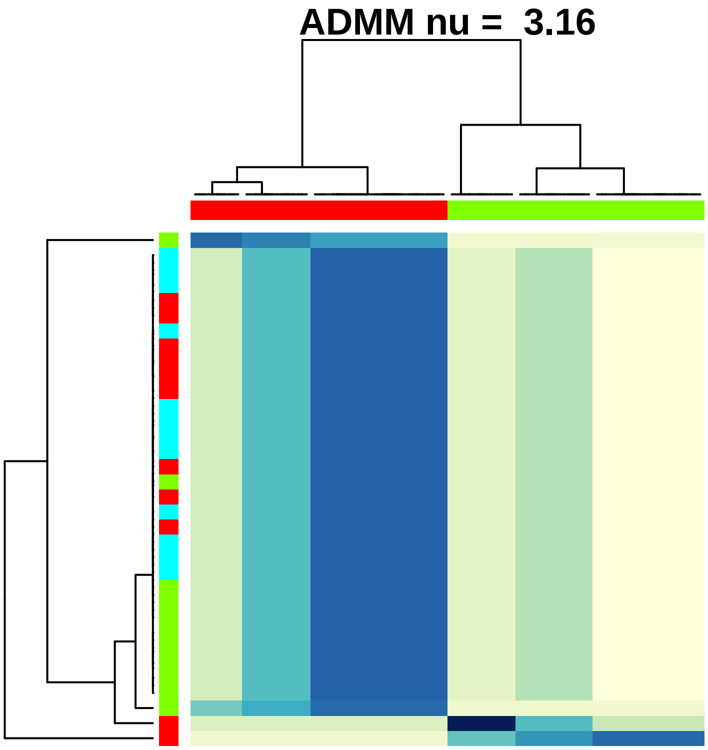


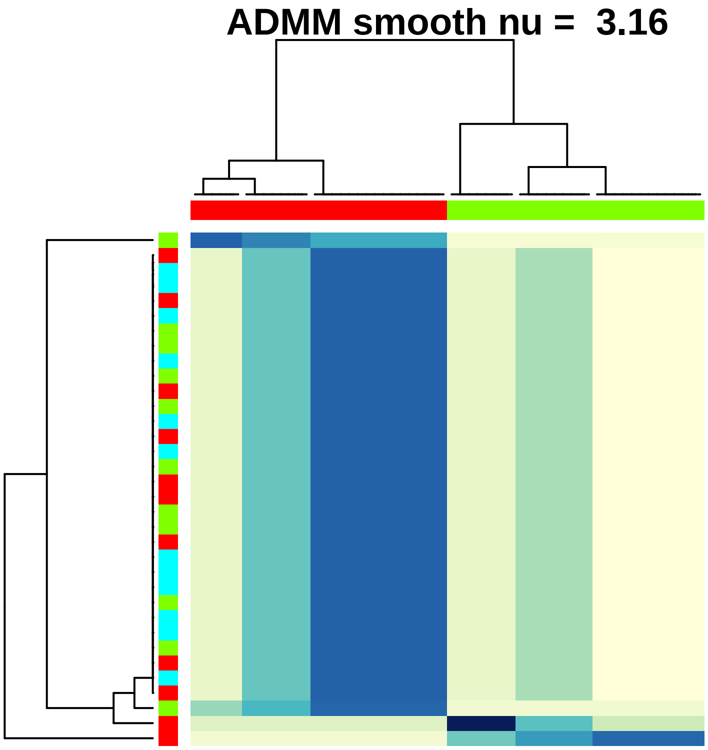
Eric's method



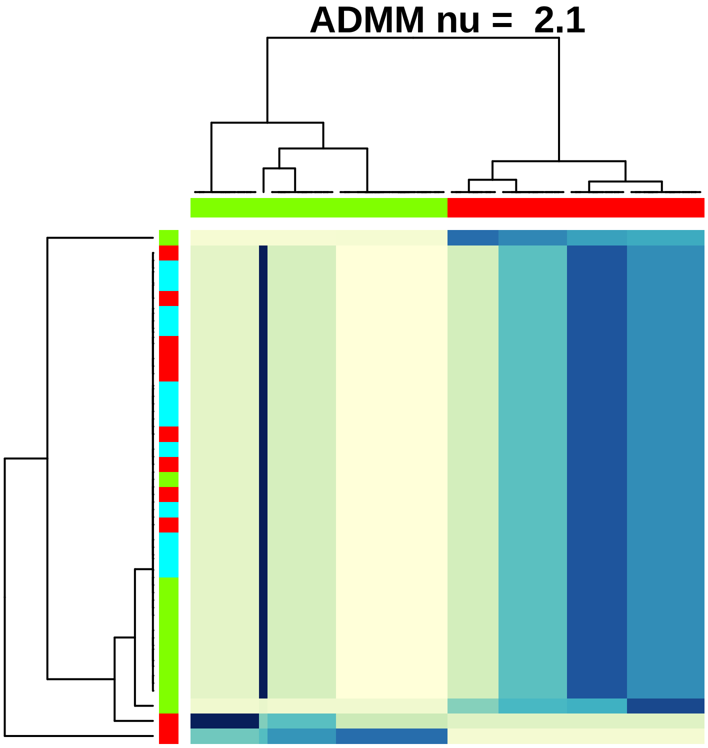
ADMM

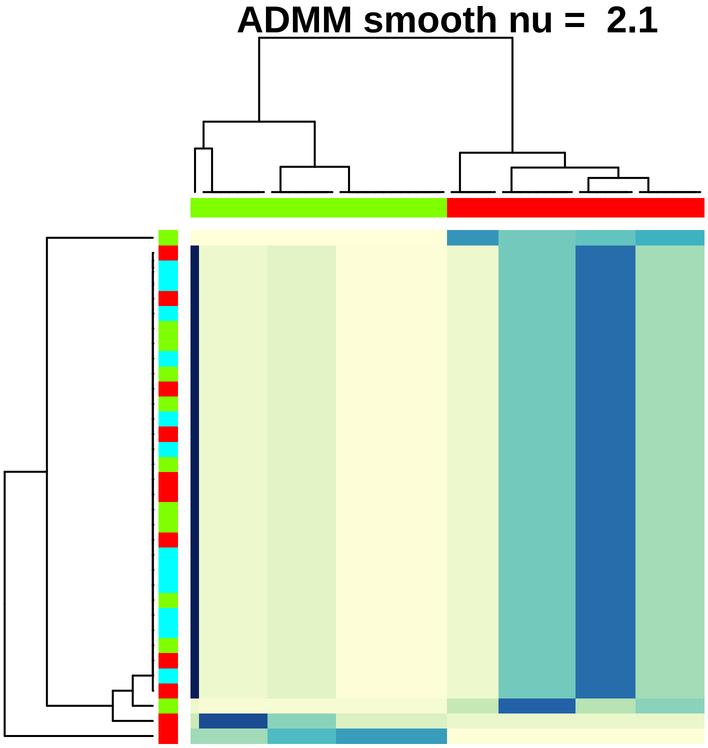
- ADMM with the same parameter (nu1, nu2) with Eric's method





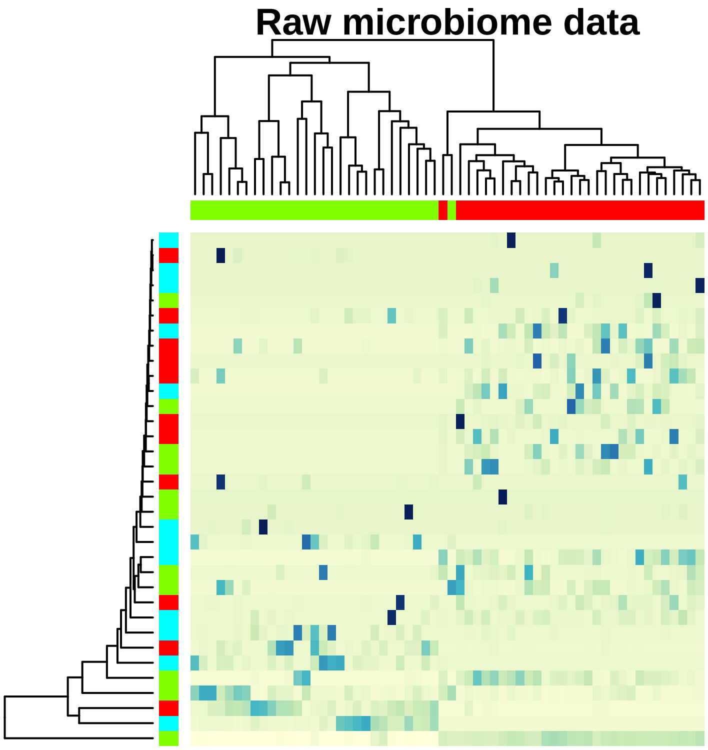
- try another parameter



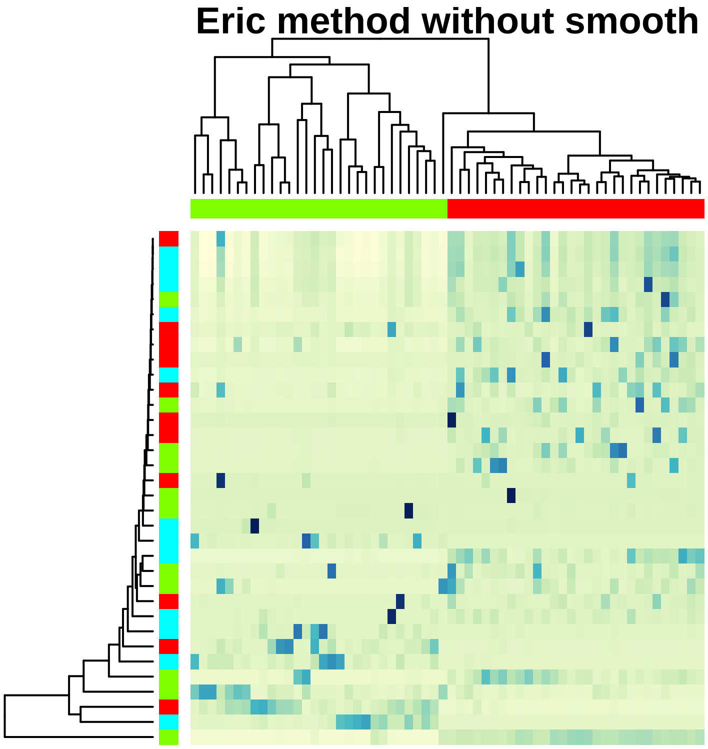


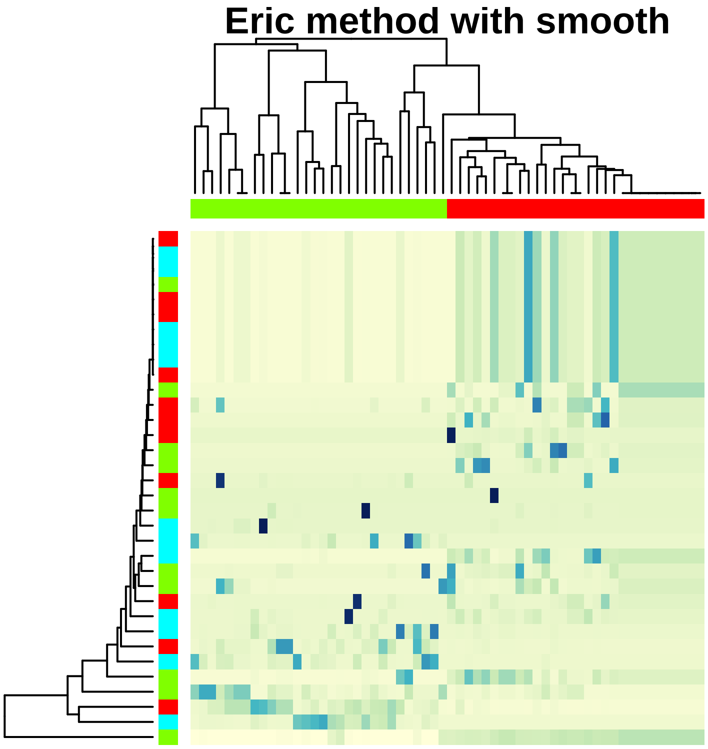
Scenario 3

Heatmap of raw data


Raw microbiome data

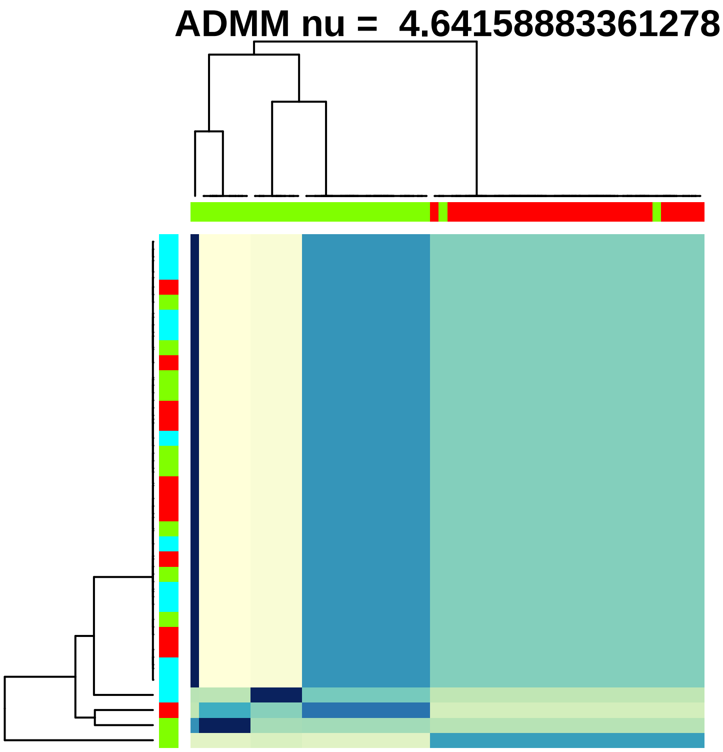
Eric's method

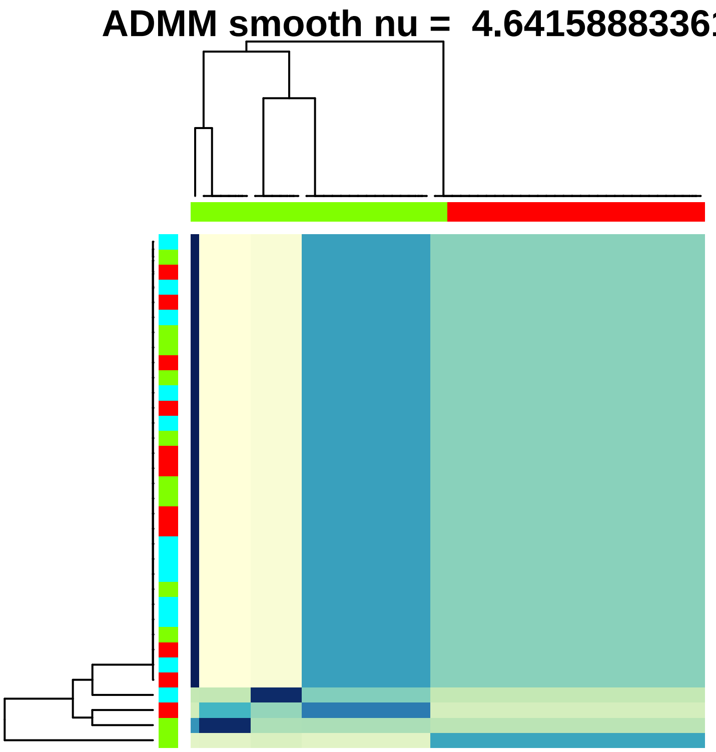

Eric method without smooth


Eric method with smooth

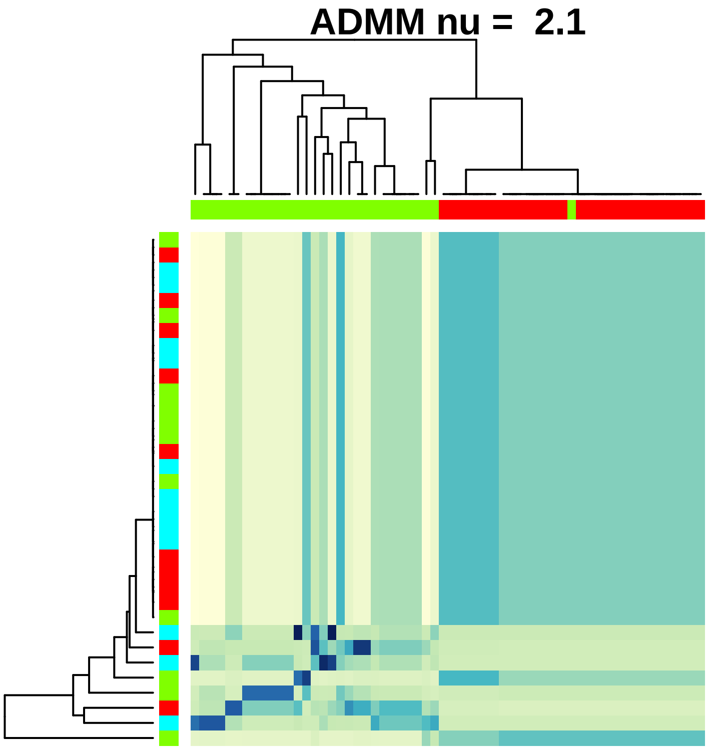
ADMM

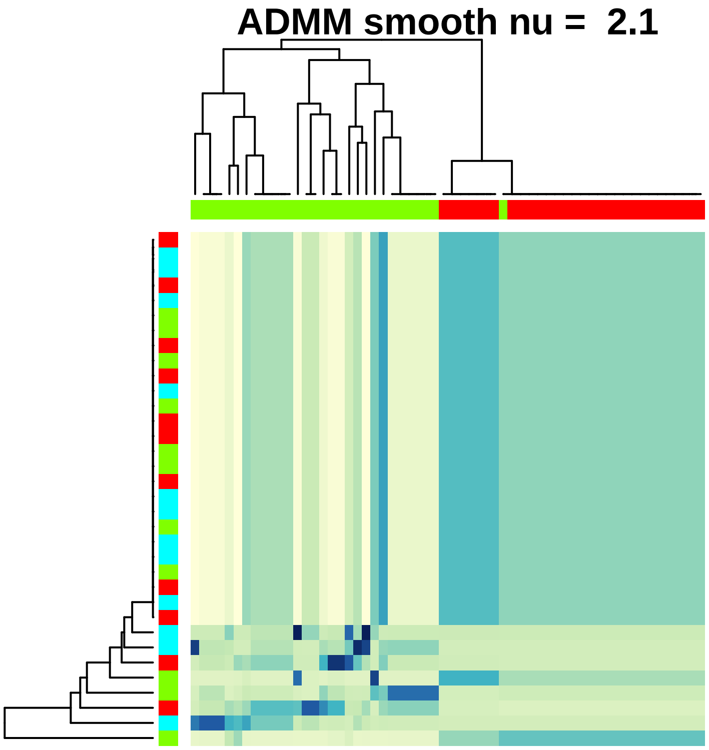
- ADMM with the same parameter (nu1, nu2) with Eric's method





- try another parameter





4. Results of the simulated microbiome data with two clusters

Devide microbiome into two groups. I tried 3 scenarios.

- scenario 1: value in [control] and [group 1] times 2
- scenario 2: value in [control] and [group 1] times 20.
- scenario 3: value in [control] and [group 1] times 200

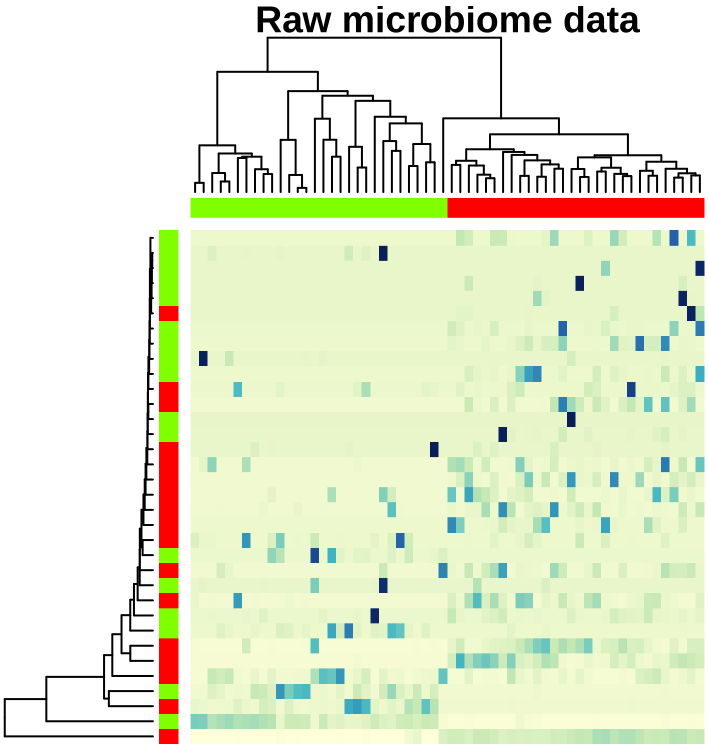
Scenario 1:

Zero percentage:

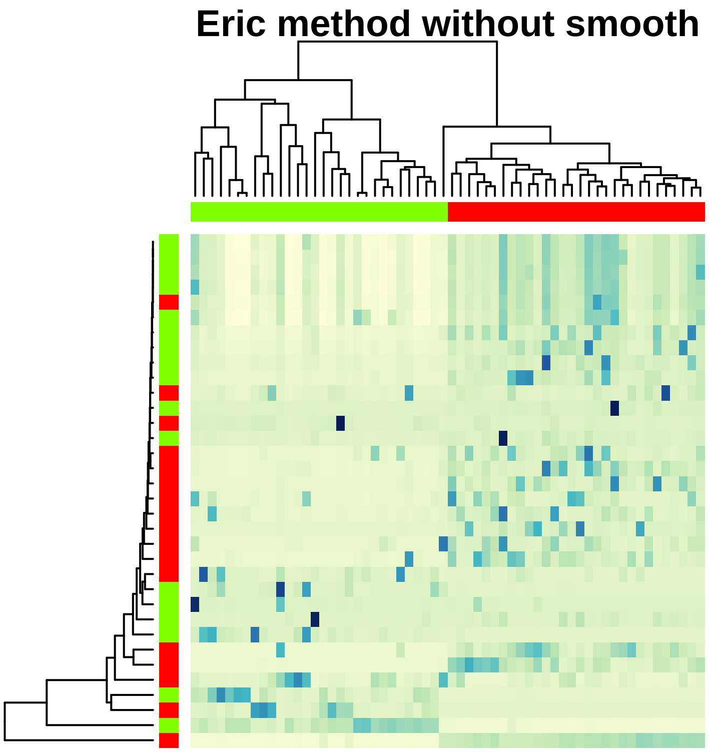
```
sum(abs(microbiome4)<0.001)/(34*60)
```

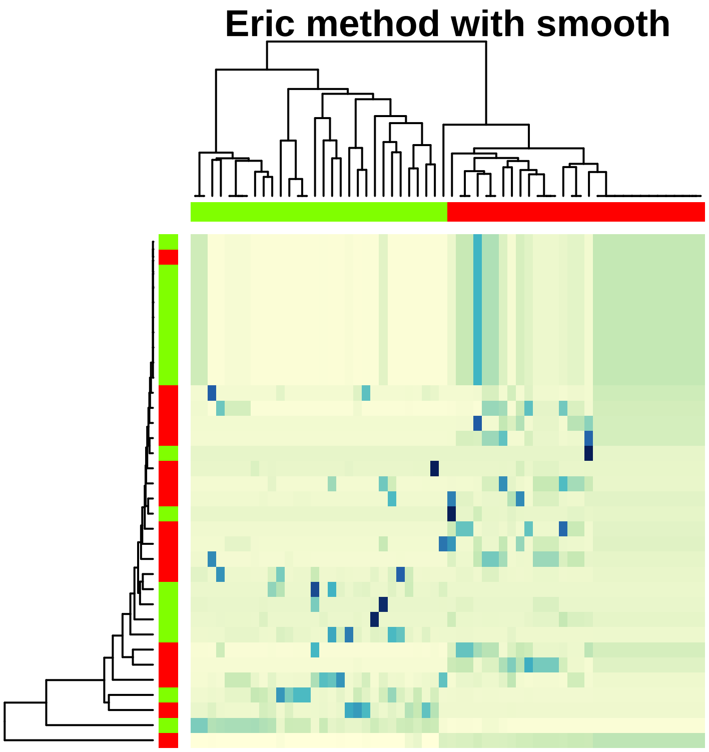
```
## [1] 0.6568627
```

Heatmap of raw data



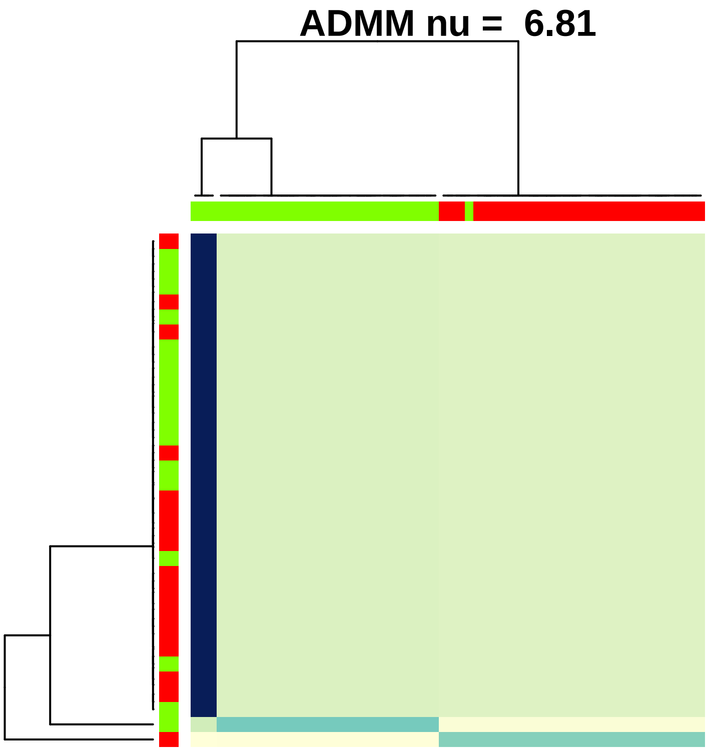
Eric's method

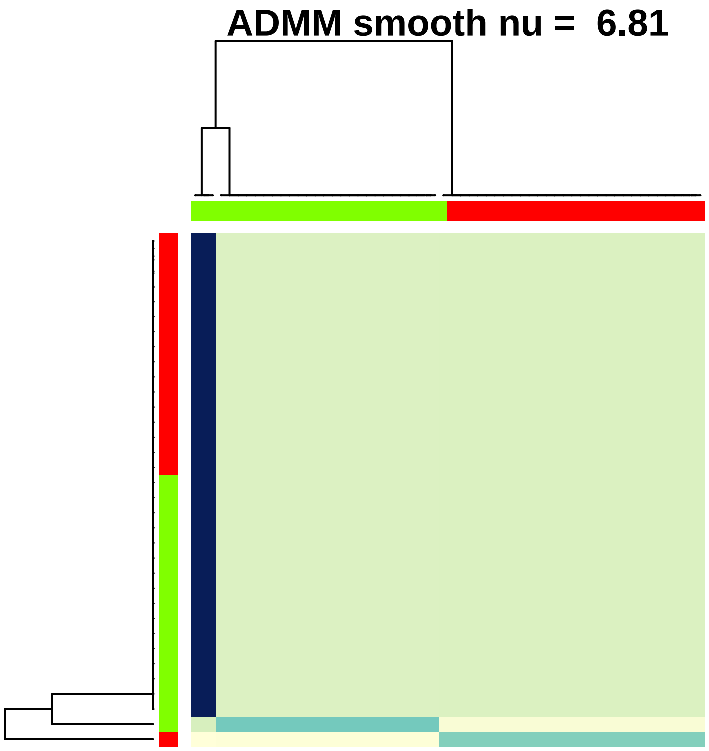

Eric method without smooth


Eric method with smooth

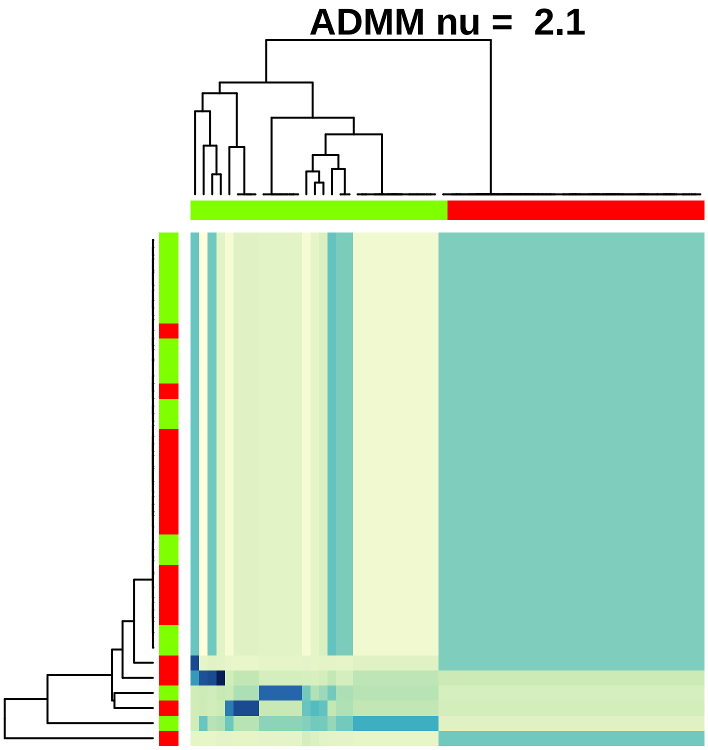
ADMM

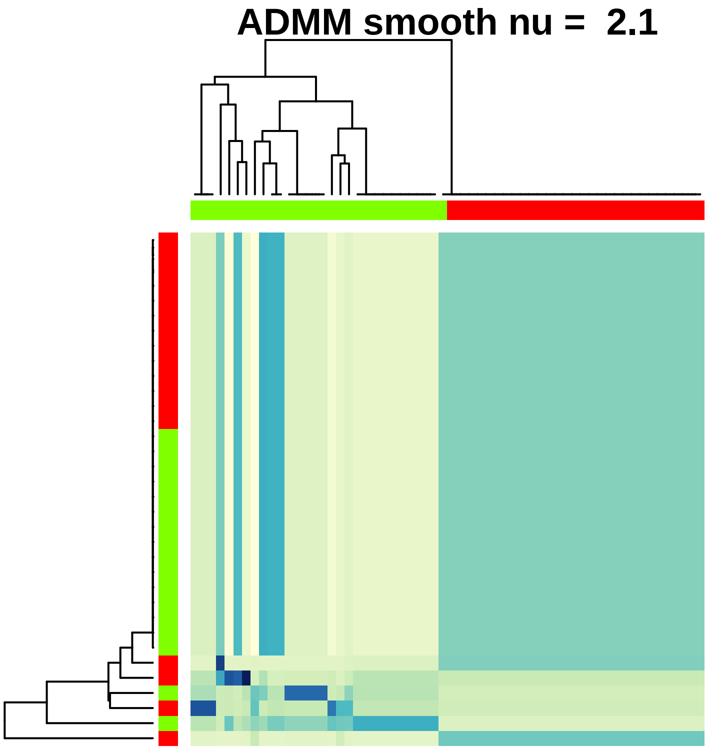
- ADMM with the same parameter (nu1, nu2) with Eric's method


ADMM nu = 6.81


ADMM smooth nu = 6.81

- try another parameter





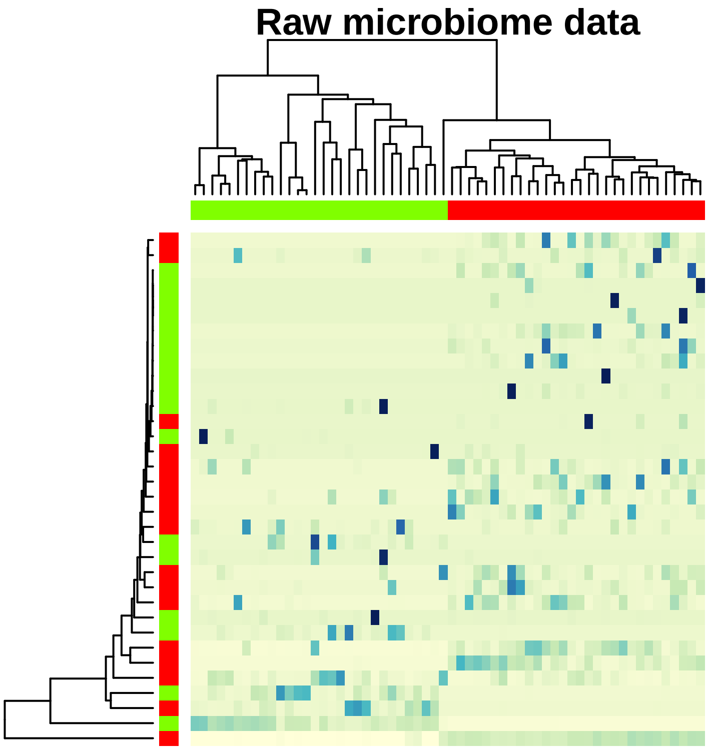
Scenario 2:

Zero percentage:

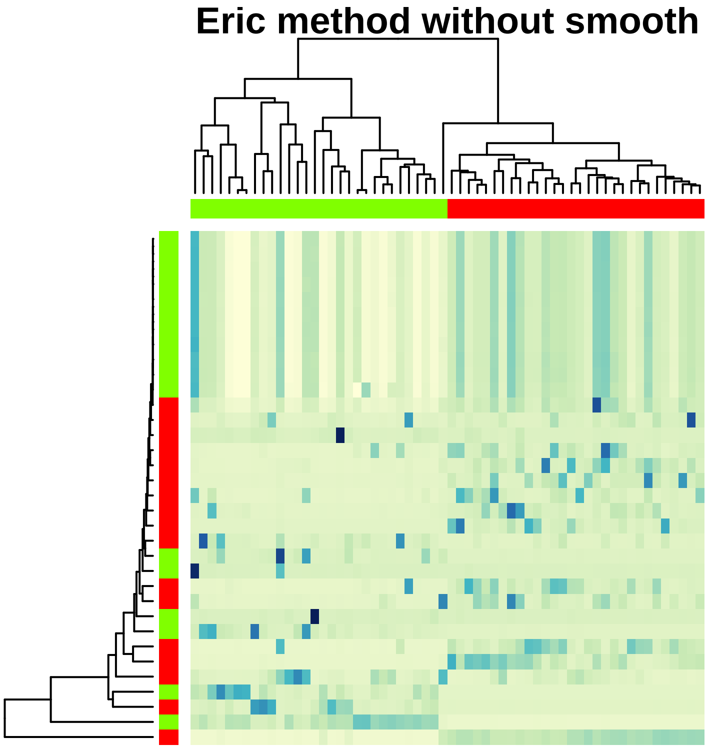
```
sum(abs(microbiome4)<0.001)/(34*60)
```

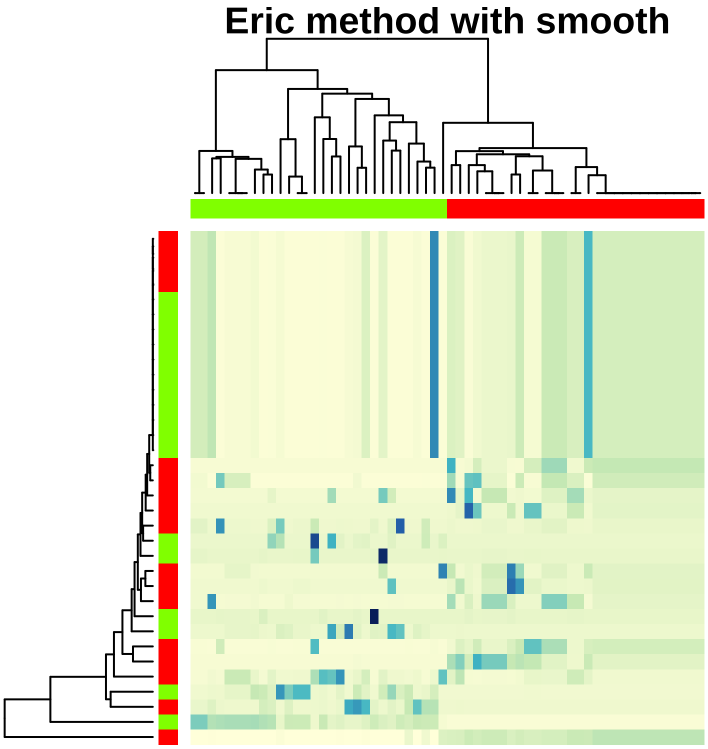
```
## [1] 0.6970588
```

Heatmap of raw data



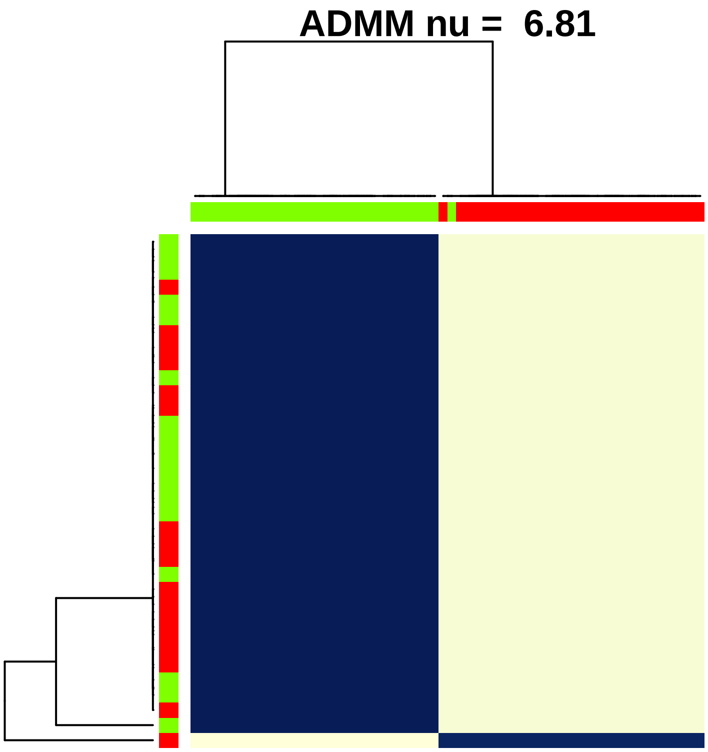
Eric's method

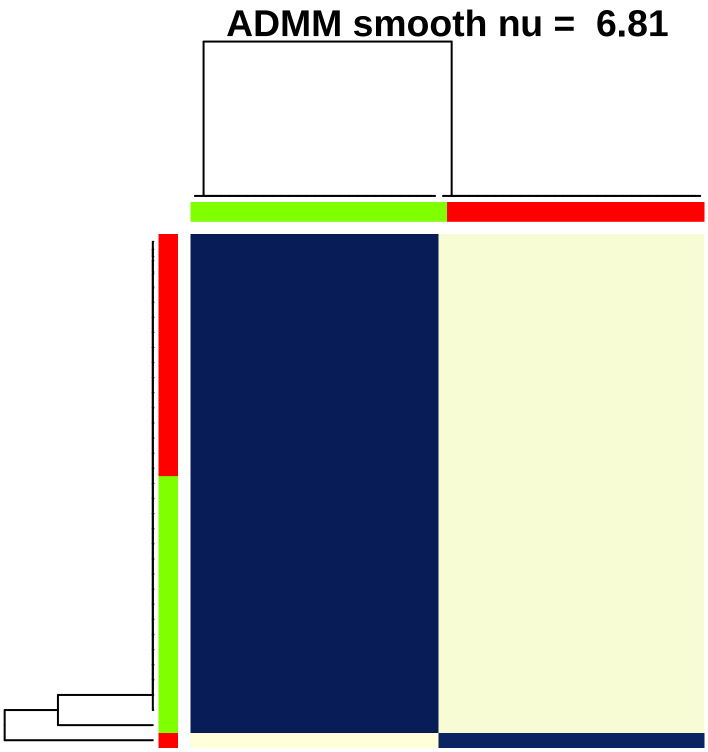




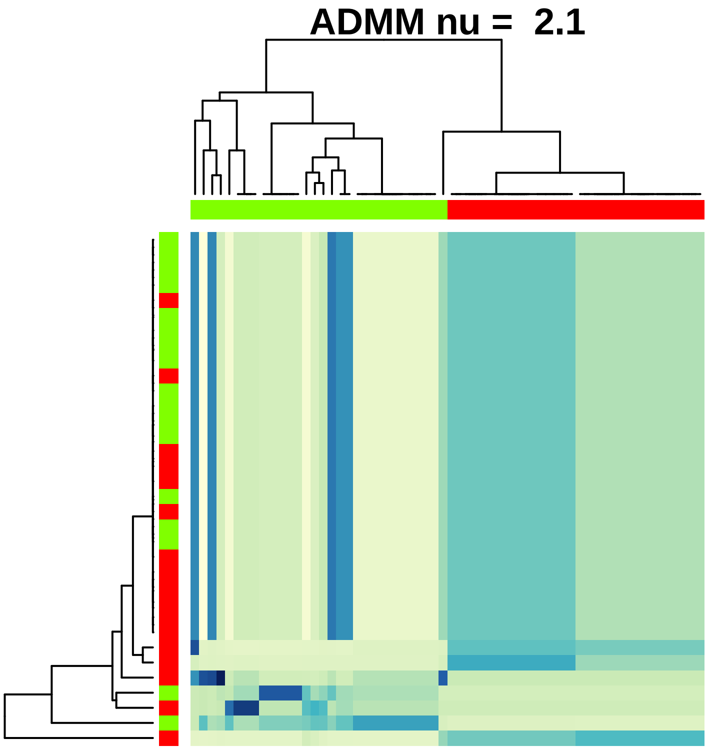
ADMM

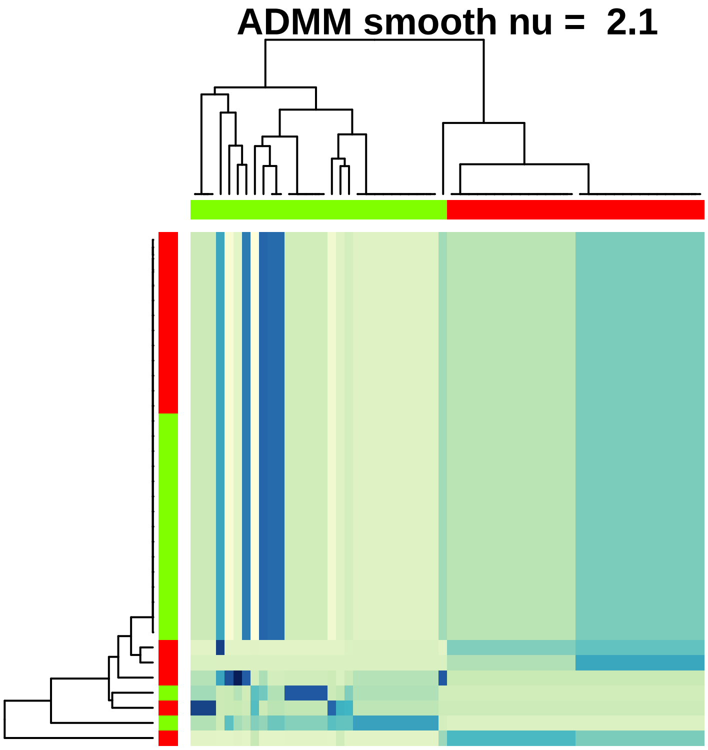
- ADMM with the same parameter (nu1, nu2) with Eric's method





- try another parameter





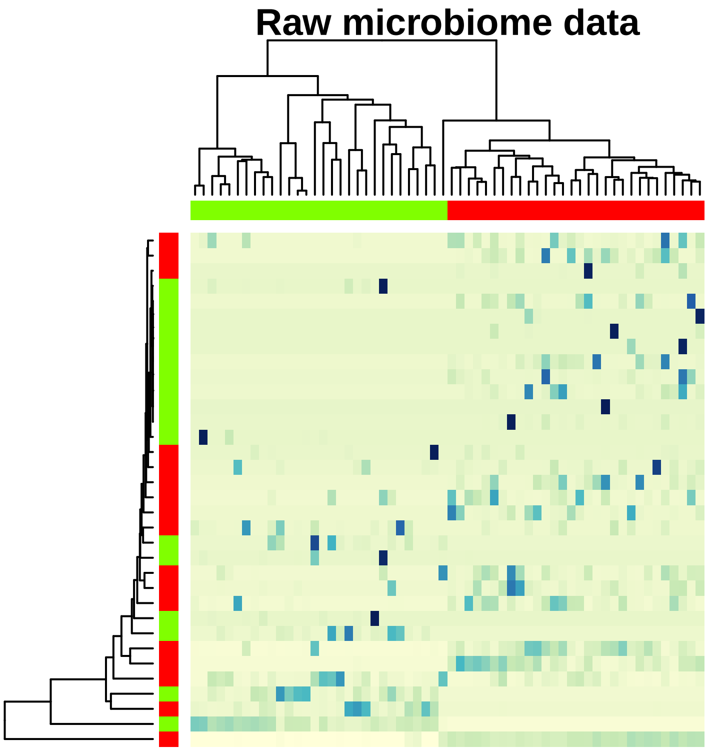
Scenario 3:

Zero percentage:

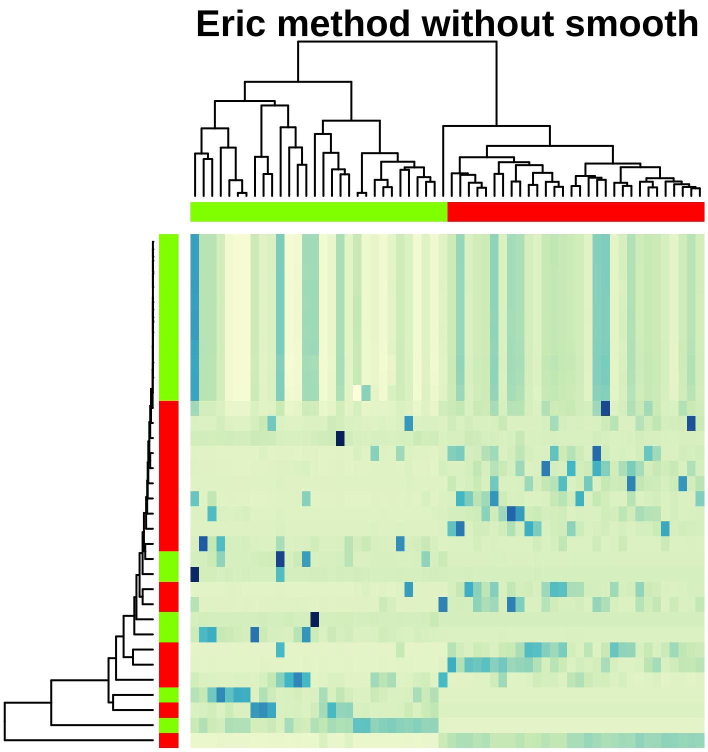
```
sum(abs(microbiome4)<0.001)/(34*60)
```

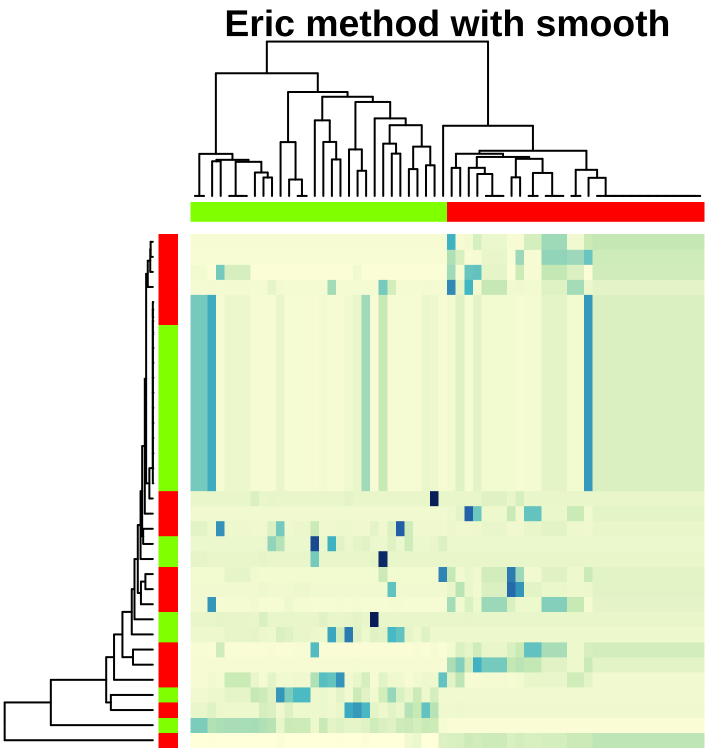
```
## [1] 0.722549
```

Heatmap of raw data



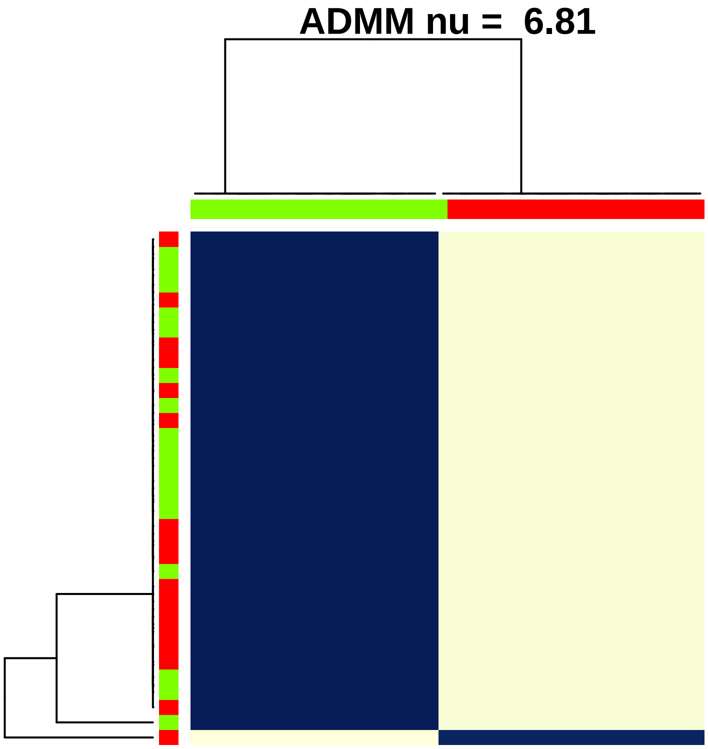
Eric's method

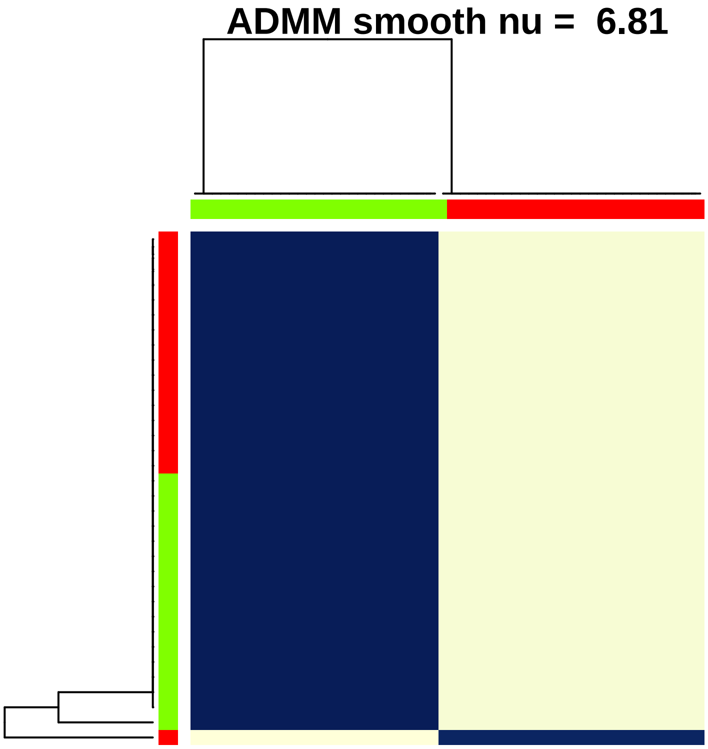




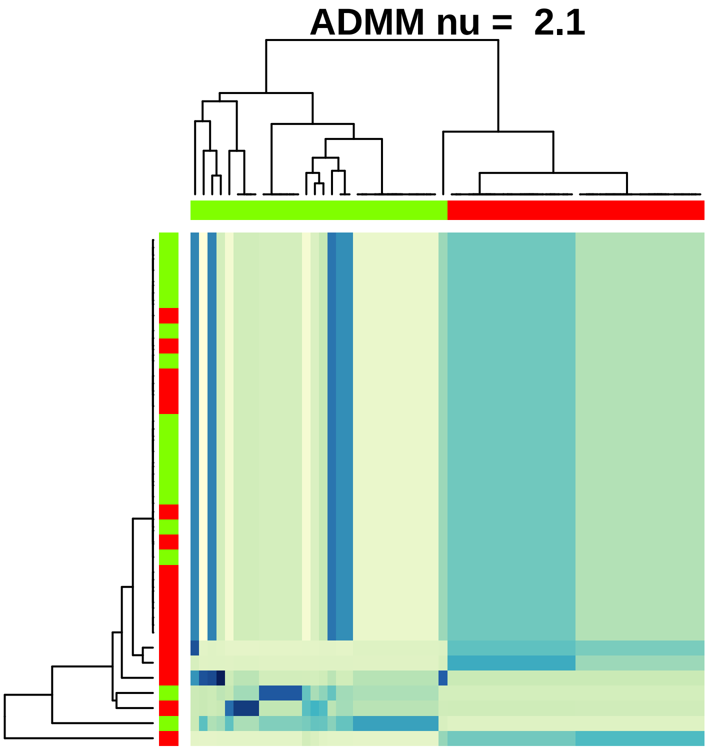
ADMM

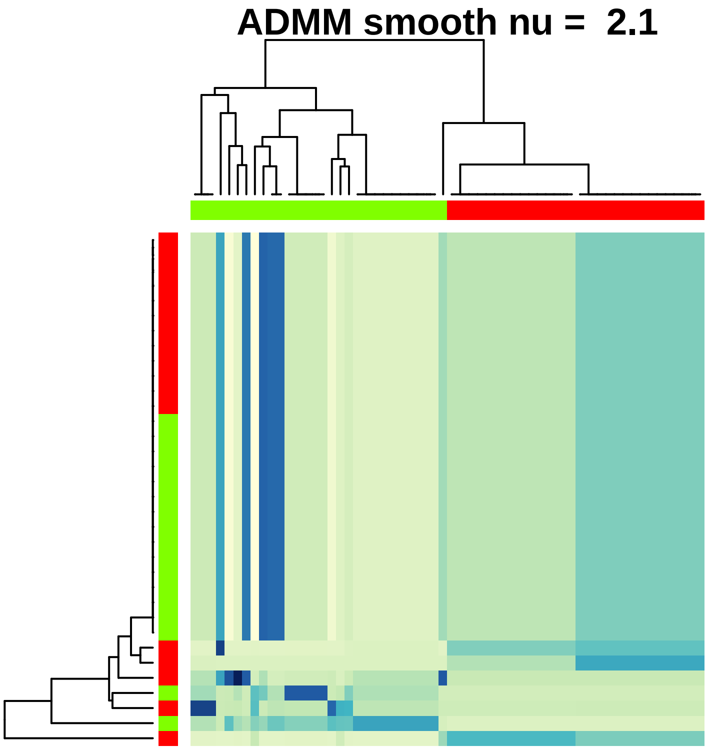
- ADMM with the same parameter (nu1, nu2) with Eric's method





- try another parameter





Summary

The ADMM can get similar results with Eric's method and is better than Eric's method some times

It does not perform well when there is three groups in the microbiome. It is easier to classify two groups than three groups.

The result with production of 200 looks better than the one with 20. The one with 20 looks better than the one with 2. The bigger the differences are, the easier to classify.

Does the real data results look ok?