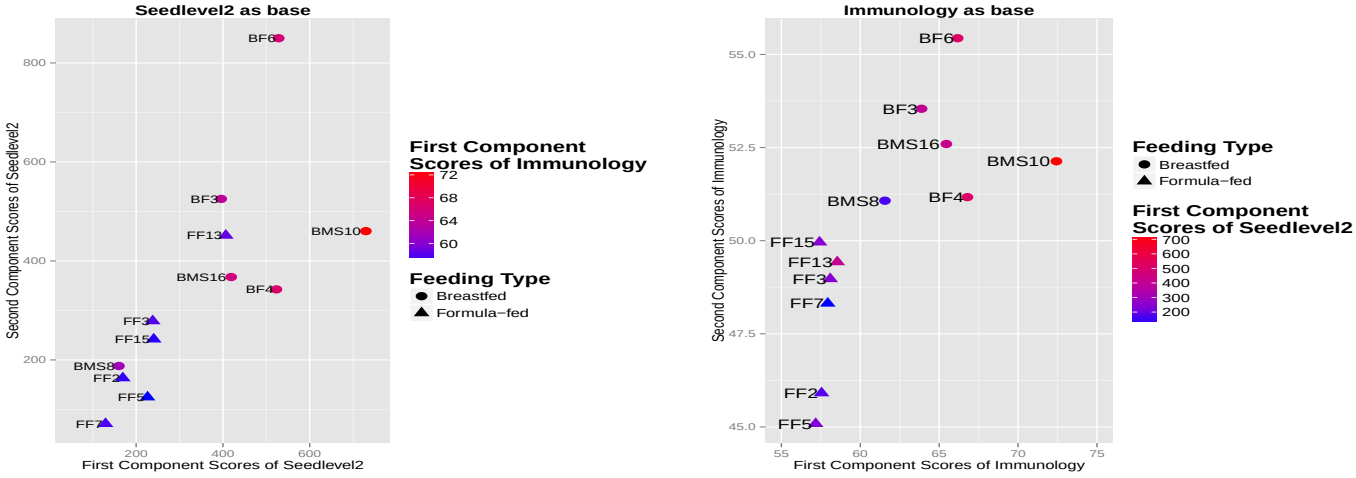
Work Summary and Prospective Plan

Kejun He

From Oct to the end of November 2014, I was working on reproducing the statistical graphs and results from *A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response*, authored by Scott Schwartz. This part of the work processes pretty well and it has been almost done. I am grateful to Jason and Dr. Ivanov for their generous help and constructive suggestions.

However, since the Scott's gene list of *Intestinal* is missing and some variables are not defined in the script, some results can not be totally reproduced. In addition, there is no "`set.seed()`" for the generation of randomness in the script which makes some reproduced plots look different from the paper's. Moreover, the CCA part of Scott's paper did not use sparse structure, thus the gene selection is questionable. Therefore, I turn to the direction based on sparse CCA, and the tuning parameter sparsity is chosen with most significant value from permutation. As a result, my gene selection from *Immunology* consists of `CD40, SEMA4D, TACR1, HLA-DOB, CCL22, CD9, GSTM4, CCL18, LRRC8E, SNED1` based on the first component score, which is different from Scott's gene selection list on his page 8. And the 5 most "correlated" microbial species include "`Virulence - Resistance to antibiotics and toxic compounds, Virulence - Virulence, Cell Wall and Capsule - Capsular and extracellular polysacchrides, Respiration - Electron donating reactions, Carbohydrates - Aminosugars`"



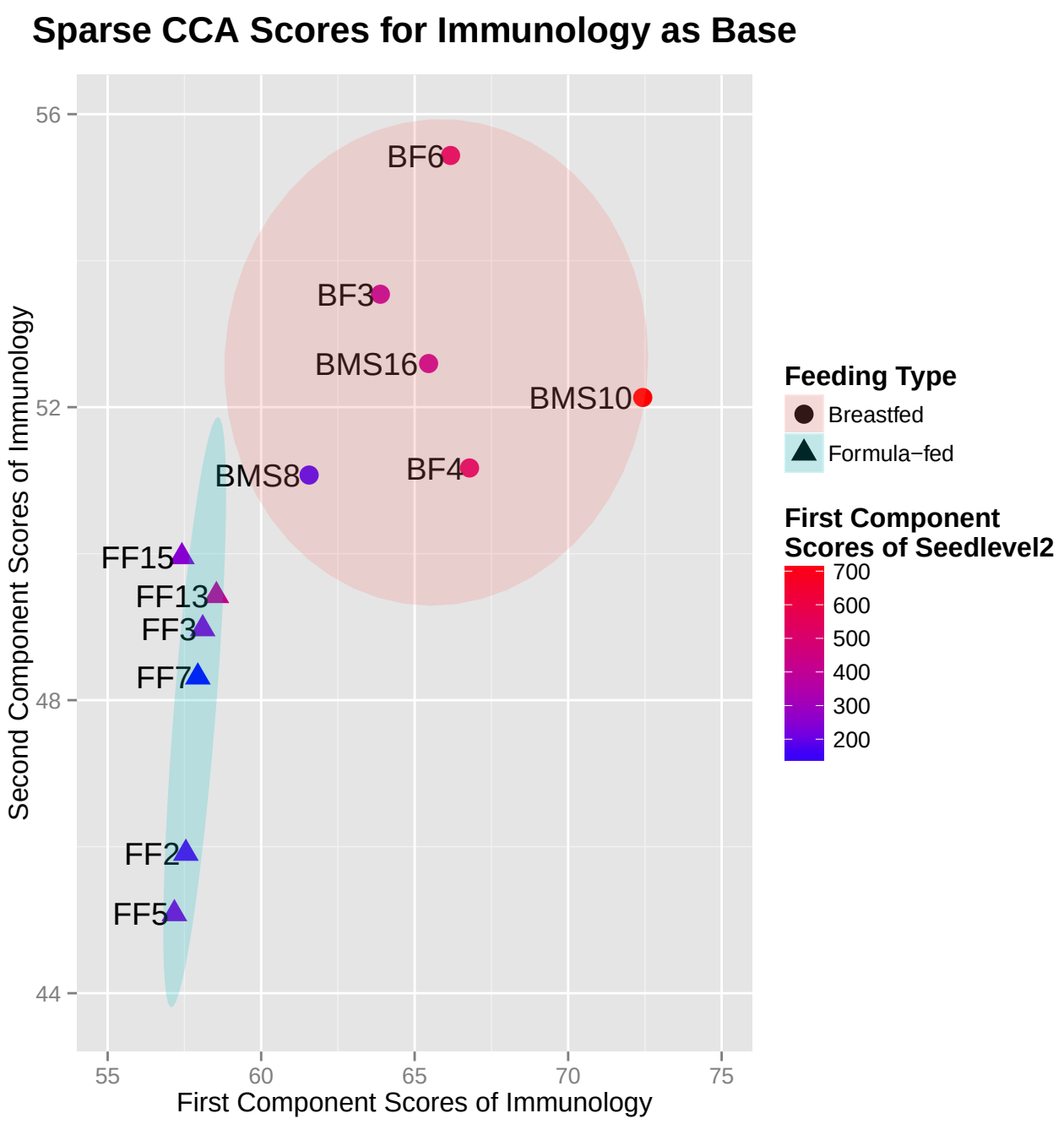
The first plot above shows each subject on the base of *first and second components scores of Seedlevel2* microbial community with different colors describing the first component scores of *Immunology* and different shapes based on the difference of feeding types. And the second plot above shows the difference when we select the top 5 Seedlevel species as well.

In future, I am gonna work with Dr. Qian to improve the CCA methodology, and our main approach may take into account the group structures, such as graph Laplacian, and proportional data constraints.

```
## redo the sparse CCA with permuted best penalties choises.
library(PMA)
cca_permute_both<-CCA.permute(x=cca_seed2,z=cca_microarray_subjects,
                              typex="standard",typez="standard",nperms=30)
```

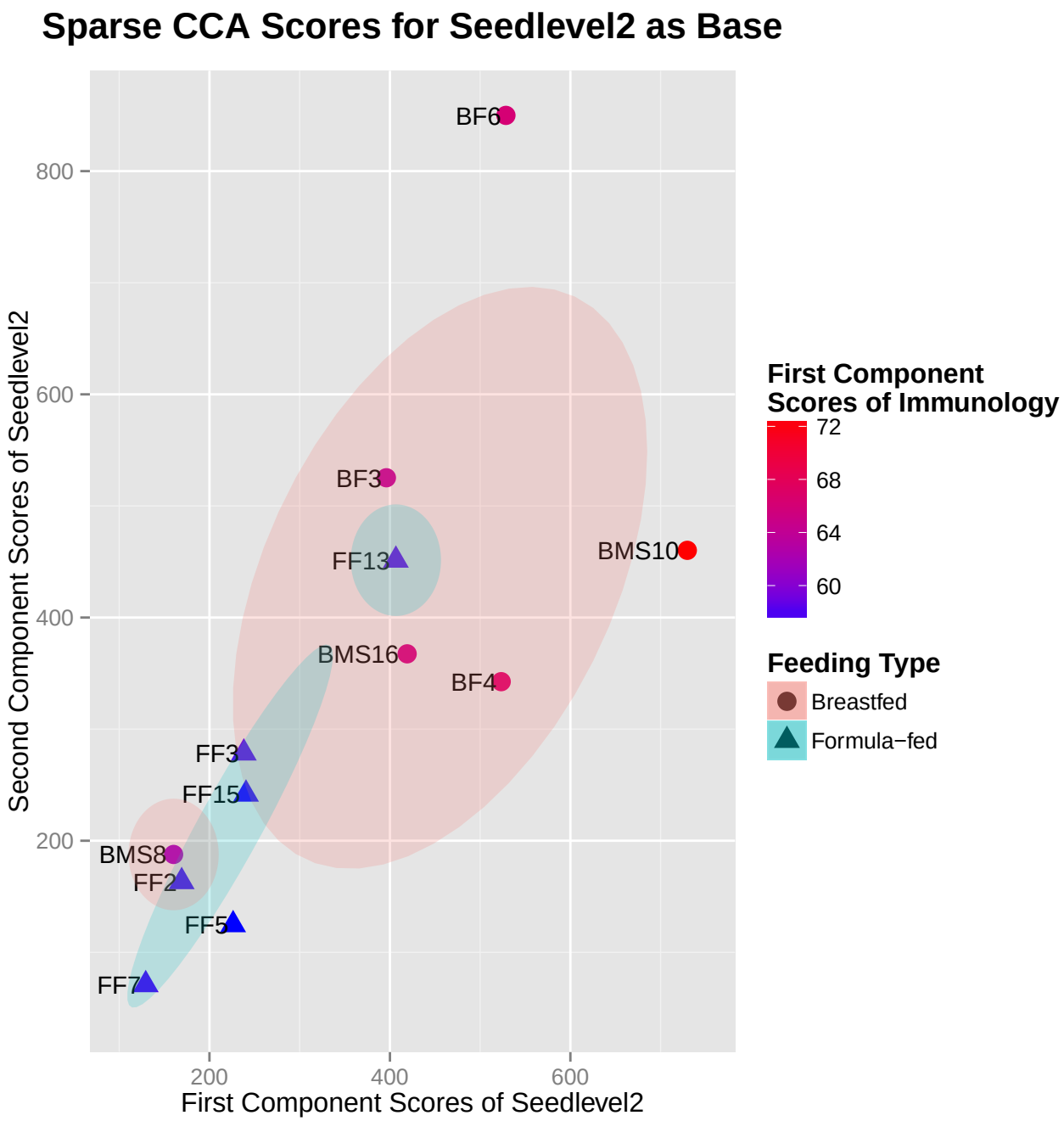
```
pen_x<-cca_permute_both$bestpenaltyx
pen_z<-cca_permute_both$bestpenaltyz
cca_out_both_2<-CCA(x=cca_seed2,z=cca_microarray_subjects,
                    typex="standard",typez="standard",
                    penaltyx=pen_x,penaltyz=pen_z,
                    xnames=abbreviate(colnames(cca_seed2),min=20),
                   znames=colnames(cca_microarray_subjects),
                   K=2)

## 12345678910111213141 5
## 12345678910111213141 5

#cca_out_both£u
cca_scores_u_2<-cca_seed2%*%cca_out_both_2$u
cca_scores_v_2<-cca_microarray_subjects%*%cca_out_both_2$v
cca_scores_2<-cbind(cca_scores_u_2,cca_scores_v_2)
colnames(cca_scores_2)<-c("U1","U2","V1","V2")
cca_scores_2<-as.data.frame(cca_scores_2)
cca_scores_2$type<-c(rep("BF",6),rep("FF",6))

library(ggplot2)
myplot2_2<-ggplot(cca_scores_2,aes(x=V1,y=V2,col=U1,shape=type))
myplot2_2<-myplot2_2+geom_point(size=4)
myplot2_2<-myplot2_2+scale_colour_continuous(name="First Component\nScores of Seedlevel2",
                                             low="blue",high="red")
myplot2_2<-myplot2_2+scale_shape_discrete(name="Feeding Type",
                                          labels=c("Breastfed","Formula-fed"))
myplot2_2<-myplot2_2+geom_text(aes(label=rownames(cca_scores_2)),col="black",hjust=1.1,size=5
myplot2_2<-myplot2_2+scale_x_continuous("First Component Scores of Immunology",
                                        limits=c(55,75))
myplot2_2<-myplot2_2+scale_y_continuous("Second Component Scores of Immunology")
myplot2_2<-myplot2_2+labs(title="Sparse CCA Scores for Immunology as Base")
myplot2_2<-myplot2_2+theme(legend.title = element_text(size=12),
                       plot.title = element_text(size=16,vjust=2.0, face="bold"),
                       legend.text=element_text(size=10))
setInternet2(T) ## to make https work
source("https://raw.githubusercontent.com/low-decarie/FAAV/master/r/stat-ellipse.R")
myplot2_2<-myplot2_2+stat_ellipse(aes(fill=type),level=0.85,alpha=0.2,
                                  geom="polygon",linetype=2)
#level for the scale of elli, alpha for the darkness of col,"polygon" for the cover of elli
#linetype=2 make the boundary transparent.
```

```
myplot2_2<-myplot2_2+scale_fill_discrete(name="Feeding Type",
                                         labels=c("Breastfed","Formula-fed"))
myplot2_2
```




```
myplot1_2<-ggplot(cca_scores_2,aes(x=U1,y=U2,col=V1,shape=type))
myplot1_2<-myplot1_2+geom_point(size=4)
myplot1_2<-myplot1_2+scale_colour_continuous(name="First Component\nScores of Immunology",
                                             low="blue",high="red")
myplot1_2<-myplot1_2+scale_shape_discrete(name="Feeding Type",
                                         labels=c("Breastfed","Formula-fed"))
```

```
myplot1_2<-myplot1_2+geom_text(aes(label=rownames(cca_scores)),col="black",hjust=1.1,size=4)
myplot1_2<-myplot1_2+scale_x_continuous("First Component Scores of Seedlevel2",
                                        limits=c(100,750))
myplot1_2<-myplot1_2+scale_y_continuous("Second Component Scores of Seedlevel2")
myplot1_2<-myplot1_2+labs(title="Sparse CCA Scores for Seedlevel2 as Base")
myplot1_2<-myplot1_2+theme(legend.title = element_text(size=12),
                      plot.title = element_text(vjust=2.0,size=16, face="bold"),
                      legend.text=element_text(size=10))
myplot1_2<-myplot1_2+stat_ellipse(aes(fill=type),level=0.6,alpha=0.2,
                                  geom="polygon",linetype=2)
cca_scores_special<-cca_scores_2[rep(10,100),]
for(t in 1:100){
 cca_scores_special[t,1:2]<-cca_scores_special[t,1:2]+c(sin(t*2*pi/100)*50,cos(t*2*pi/100)*50
}
myplot1_2<-myplot1_2+geom_polygon(aes(x=U1,y=U2,fill=type),cca_scores_special,
                                  alpha=0.2,linetype=0)
cca_scores_special2<-cca_scores_2[rep(1,100),]
for(t in 1:100){
 cca_scores_special2[t,1:2]<-cca_scores_special2[t,1:2]+c(sin(t*2*pi/100)*50,cos(t*2*pi/100)*
}
myplot1_2<-myplot1_2+geom_polygon(aes(x=U1,y=U2,fill=type),cca_scores_special2,
                                  alpha=0.2,linetype=0)
myplot1_2<-myplot1_2+scale_fill_discrete(name="Feeding Type",
                                         labels=c("Breastfed","Formula-fed"))
myplot1_2
```

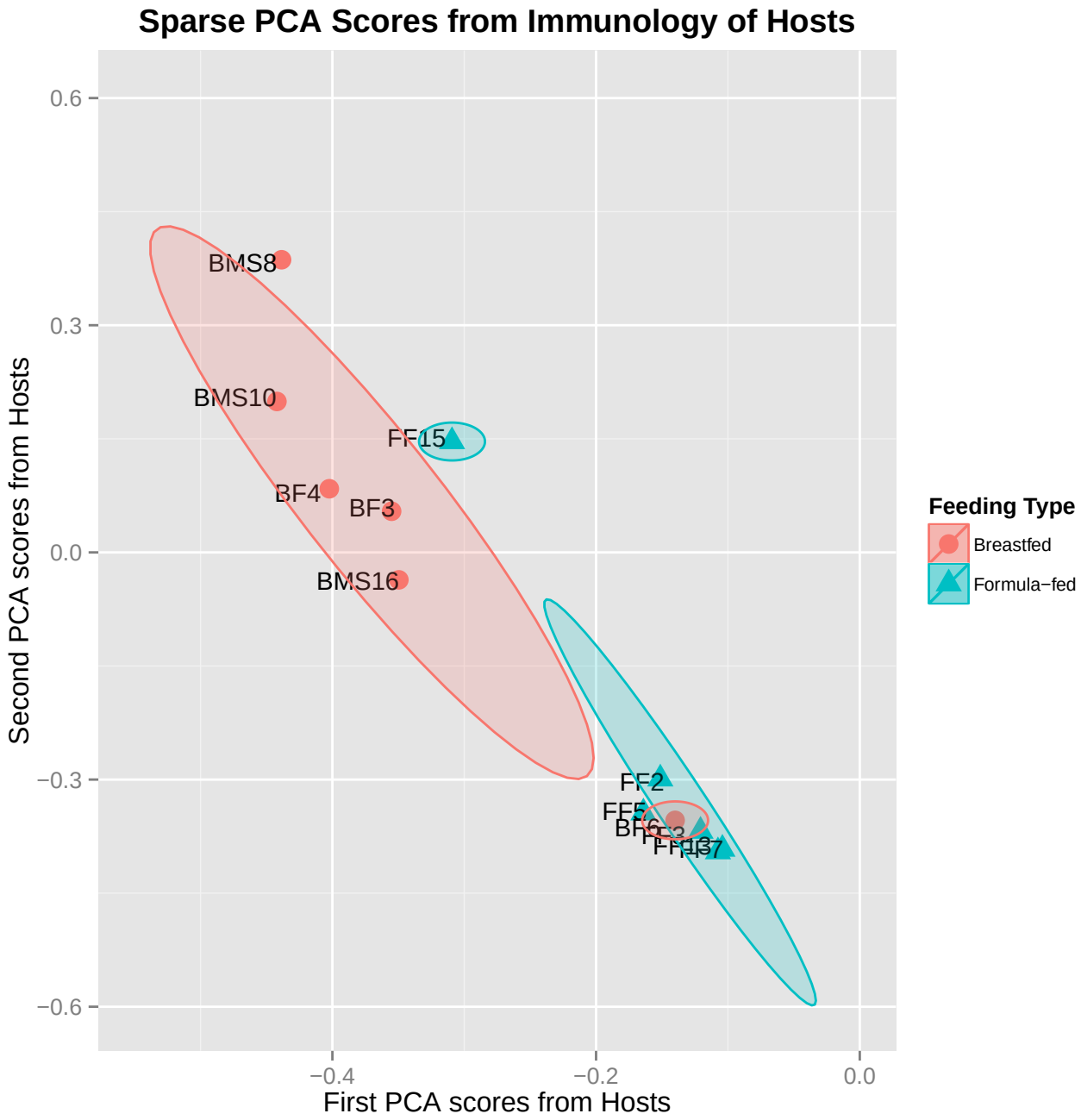



```
## re select the names
cca_seed2_t<-t(cca_seed2)
cca_seed2_t<-cca_seed2_t*cca_out_both_2$u[,1]
cca_seed2_tt<-t(cca_seed2_t)
cca_seed2_tt<-cca_seed2_tt*sign(cca_scores_u_2[,1])
cca_seed2_select_new<-apply(cca_seed2_tt,2,sum)
names(cca_seed2_select_new)<-colnames(cca_seed2)
cca_seed2_select_new_order<-sort(abs(cca_seed2_select_new),decreasing = T) # seed first comp

## re select the microarray
cca_microarray_subjects_t<-t(cca_microarray_subjects)
```

```
cca_microarray_subjects_t<-cca_microarray_subjects_t*cca_out_both_2$v[,1]
cca_microarray_subjects_tt<-t(cca_microarray_subjects_t)
cca_microarray_subjects_tt<-cca_microarray_subjects_tt*sign(cca_scores_v_2[,1])
cca_micro_select_new<-apply(cca_microarray_subjects_tt,2,sum)
names(cca_micro_select_new)<-colnames(cca_microarray_subjects)
cca_micro_select_new_order<-sort(abs(cca_micro_select_new),
                                 decreasing=T) # microarray fist compo sort
```

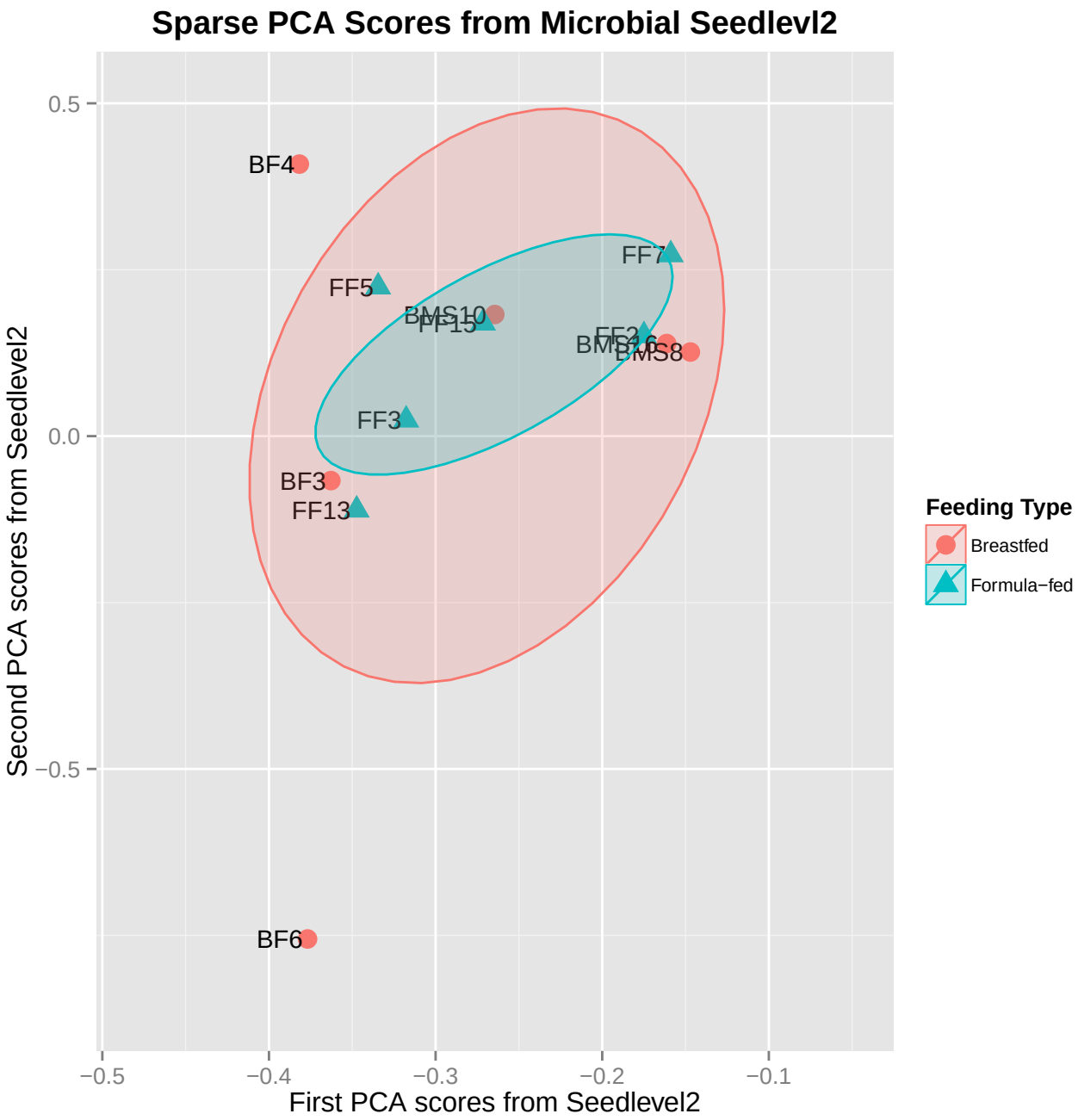
```
library(PMA)
spca_immun_cv<-SPC.cv(x=cca_microarray_subjects,sumabsvs=seq(1.2,25,by=0.2),
                      nfolds=10,niter=10)

##  Fold  1  out of  10
##  Fold  2  out of  10
##  Fold  3  out of  10
##  Fold  4  out of  10
##  Fold  5  out of  10
##  Fold  6  out of  10
##  Fold  7  out of  10
##  Fold  8  out of  10
##  Fold  9  out of  10
##  Fold  10  out of  10

spca_immun<-SPC(x=cca_microarray_subjects,sumabsv=spca_immun_cv$bestsumabsv,
                K=2)

## 12345678910111213141516171819 20
## 1234567891011

spca_scores<-spca_immun$u
rownames(spca_scores)<-rownames(cca_microarray_subjects)
spca_scores<-as.data.frame(spca_scores)
names(spca_scores)<-c("Host1","Host2")
spca_scores$type<-c(rep("BF",6),rep("FF",6))
library(ggplot2)
spca_plot1<-ggplot(spca_scores,aes(x=Host1,y=Host2,shape=type,colour=type,
                                   ymax=max(abs(Host2))*1.5))
spca_plot1<-spca_plot1+geom_point(size=4)
spca_plot1<-spca_plot1+scale_shape_discrete(name="Feeding Type",
                                            labels=c("Breastfed","Formula-fed"))
spca_plot1<-spca_plot1+scale_color_discrete(name="Feeding Type",
                                            labels=c("Breastfed","Formula-fed"))
spca_plot1<-spca_plot1+geom_text(aes(label=rownames(spca_scores)),
                                 col="black",hjust=1.1,size=4,
                                 position=position_jitter(height=0.01,width=0.01))
## the position=position_jitter(height=0.01,width=0.01) will make the material in the same
```

```r
# a litter bit separate.
## wihle position_dodge will separate the material between groups.

spca_plot1<-spca_plot1+scale_x_continuous("First PCA scores from Hosts",
                                          limits=c(-0.55,0.0))
spca_plot1<-spca_plot1+scale_y_continuous("Second PCA scores from Hosts")
spca_plot1<-spca_plot1+labs(title="Sparse PCA Scores from Immunology of Hosts")
spca_plot1<-spca_plot1+theme(legend.title=element_text(size=10),
                             legend.text=element_text(size=8),
                             plot.title=element_text(size=14,vjust=1.1,face="bold"))
#source("https://raw.github.com/low-decarie/FAAV/master/r/stat-ellipse.R")
spca_plot1<-spca_plot1+stat_ellipse(aes(x=Host1,y=Host2,fill=type),
                                    level=0.8,geom="polygon",alpha=0.2)
spca_plot1<-spca_plot1+scale_fill_discrete(name="Feeding Type",
                                           labels=c("Breastfed","Formula-fed"))
#level for the scale of elli, alpha for the darkness of col,"polygon" for the cover of elli
spca_special<-spca_scores[rep(6,100),]
for(t in 1:100){
 spca_special[t,1:2]<-spca_special[t,1:2]+c(sin(t*2*pi/100)/40,cos(t*2*pi/100)/40)
}
spca_plot1<-spca_plot1+geom_polygon(aes(x=Host1,y=Host2,fill=type),spca_special,
                                    alpha=0.2)
spca_special2<-spca_scores[rep(11,100),]
for(t in 1:100){
 spca_special2[t,1:2]<-spca_special2[t,1:2]+c(sin(t*2*pi/100)/40,cos(t*2*pi/100)/40)
}
spca_plot1<-spca_plot1+geom_polygon(aes(x=Host1,y=Host2,fill=type),spca_special2,
                                    alpha=0.2)
spca_plot1
```




```
## sparse pca when microbial as base
spca_seed_cv<-SPC.cv(x=cca_seed2,sumabsvs=seq(1.2,12,by=0.2),
                     nfolds=10,niter=10)

##  Fold  1  out of  10
##  Fold  2  out of  10
##  Fold  3  out of  10
##  Fold  4  out of  10
##  Fold  5  out of  10
##  Fold  6  out of  10
##  Fold  7  out of  10
```

```
##  Fold  8  out of  10
##  Fold  9  out of  10
##  Fold  10  out of  10

spca_seed<-SPC(x=cca_seed2,sumabsv=spca_seed_cv$bestsumabsv,
               K=2)

## 1234567
## 12345678910111213

spca_scores_seed<-spca_seed$u
rownames(spca_scores_seed)<-rownames(cca_seed2)
spca_scores_seed<-as.data.frame(spca_scores_seed)
names(spca_scores_seed)<-c("microb1","microb2")
spca_scores_seed$type<-c(rep("BF",6),rep("FF",6))

#ggplot of Sparse pca of microbial
spca_plot2<-ggplot(spca_scores_seed,
                   aes(x=microb1,y=microb2,shape=type,colour=type))
spca_plot2<-spca_plot2+geom_point(size=4)
spca_plot2<-spca_plot2+scale_shape_discrete(name="Feeding Type",
                                            labels=c("Breastfed","Formula-fed"))
spca_plot2<-spca_plot2+scale_color_discrete(name="Feeding Type",
                                            labels=c("Breastfed","Formula-fed"))
spca_plot2<-spca_plot2+geom_text(aes(label=rownames(spca_scores_seed)),
                                 col="black",hjust=1.1,size=4)
spca_plot2<-spca_plot2+scale_x_continuous("First PCA scores from Seedlevel2",
                                          limits=c(min(spca_scores_seed[,1])-0.1,
                                                   max(spca_scores_seed[,1])+0.1))
spca_plot2<-spca_plot2+scale_y_continuous("Second PCA scores from Seedlevel2",
                                          limits=c(min(spca_scores_seed[,2])-0.1,
                                                   max(spca_scores_seed[,2])+0.1))
spca_plot2<-spca_plot2+labs(title="Sparse PCA Scores from Microbial Seedlevl2")
spca_plot2<-spca_plot2+theme(legend.title=element_text(size=10),
                             legend.text=element_text(size=8),
                             plot.title=element_text(size=14,vjust=1.1,face="bold"))
## stat_ellipse has been sourced.
spca_plot2<-spca_plot2+stat_ellipse(aes(x=microb1,y=microb2,fill=type),
                                    level=0.6,geom="polygon",alpha=0.2)
spca_plot2<-spca_plot2+scale_fill_discrete(name="Feeding Type",
                                           labels=c("Breastfed","Formula-fed"))
#level for the scale of elli, alpha for the darkness of col,"polygon" for the cover of elli
spca_plot2
```


Sparse PCA Scores from Microbial Seedlevl2
Second PCA scores from Seedlevel2
First PCA scores from Seedlevel2
0.5
0.0
−0.5
−0.5
−0.4
−0.3
−0.2
−0.1
BF4
FF7
FF5
BMS10
FF15
FF2
BMS6
BMS8
FF3
BF3
FF13
BF6
Feeding Type
Breastfed
Formula−fed