Figures in R for BSF Metagenome

```
#Load libraries
library(tidyverse)
library(ComplexHeatmap)

### Stacked TPM sample matrix in excel so that tpm and sample are columns in dataframe instead. Load
dataframes.
## dfs to plot from Data Wrangling: dfSulfurTPM, dfDenit1TPM, dfDenit2TPM, dfCarbonTPM

########################## SampleBubblePlot() ##################################
SampleBubblePlot <- function(dataframe){
  # return a bubble style plot for a gene abundance per Sample and Sample Grouping
  # Inputs: dataframe = raw data frame of gene abundance data formatted with colnames: Name,
Sample, Group, tpm

  cols <- c('Name', 'Group')
  dataframe[cols] <- lapply(dataframe[cols], factor)

  #Order x-axis samples, remove points for samples with zeros
  SamplePlot <- dataframe[which(dataframe$tpm>0),] %>%
    mutate(Sample = fct_relevel(Sample,
                              "D35.1","D33.1","D29.1","D56.1","D46.1","D12B.1","D67B.1",
                              "D33.2","D29.2","D46.2","D12B.2","D67B.2",
                              "D35.2","D35.3","D56.2","D56.3")) %>%
    ggplot(aes(x=Sample, y=Name))

  ##Plot ggplot2 object, size bubbles is tpm, color is gene
  BubblePlot <- SamplePlot +
    geom_count(aes(size=tpm, color=Name)) +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
          panel.background = element_blank(), axis.line = element_line(colour = "black")) +
    scale_size(range = c(1,15))

  return(BubblePlot)

}
SampleNitPathBubblePlot <- function(dataframe){
  # return a bubble style plot for a gene abundance per Sample and Sample Grouping
  # Inputs: dataframe = raw data frame of gene abundance data formatted with colnames: Name,
Sample, Group, tpm

  cols <- c('Name', 'Group')
  dataframe[cols] <- lapply(dataframe[cols], factor)

  #Order x-axis samples, remove points for samples with zeros
  SamplePlot <- dataframe[which(dataframe$tpm>0),] %>%
    mutate(Sample = fct_relevel(Sample,
                               "D35.1","D33.1","D29.1","D56.1","D46.1","D12B.1","D67B.1",
                               "D33.2","D29.2","D46.2","D12B.2","D67B.2",
                               "D35.2","D35.3","D56.2","D56.3"),
           Name = fct_relevel(Name,
                             "napA","napB","narI","narG","narH","nirS","norB","norC","norD",
                             "norQ","nirB","nirD","nrfA","nrfH")) %>%
    ggplot(aes(x=Sample, y=Name))

  ##Plot ggplot2 object, size bubbles is tpm, color is gene
  BubblePlot <- SamplePlot +
    geom_count(aes(size=tpm, color=Pathway)) +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
          panel.background = element_blank(), axis.line = element_line(colour = "black")) +
    scale_size(range = c(1,15))

  return(BubblePlot)

}
SampleCPathBubblePlot <- function(dataframe){
  # return a bubble style plot for a gene abundance per Sample and Sample Grouping
  # Inputs: dataframe = raw data frame of gene abundance data formatted with colnames: Name,
Sample, Group, tpm

  cols <- c('Name', 'Group')
  dataframe[cols] <- lapply(dataframe[cols], factor)

  ##Make $Pathway column

  rbcList <- c('rbcS','rbcL')
  methList <- c('mcrA','mcrB','mcrG')
  aeroCODHList <- c('coxS','coxM','coxL')
  anaeroCODHList <- c('cooS_acsA','cdhE_acsC','cdhD_acsD','cdhC','cdhA','acsB')

  for (row in 1:nrow(dataframe)) {
    if (dataframe$Name[row] %in% rbcList) {
      dataframe$Pathway[row] <- 'RuBiSCO'
    } else if (dataframe$Name[row] %in% methList) {
      dataframe$Pathway[row] <- 'Methanogenesis'
    } else if  (dataframe$Name[row] %in% aeroCODHList) {
      dataframe$Pathway[row] <- 'AerobicCODH'
    } else if  (dataframe$Name[row] %in% anaeroCODHList) {
      dataframe$Pathway[row] <- 'AnaerobicCODH'
    } else {
      print("Something is wrong")
    }}

  #Order x-axis samples, remove points for samples with zeros
  SamplePlot <- dataframe[which(dataframe$tpm>0),] %>%
    mutate(Sample = fct_relevel(Sample,
                               "D35.1","D33.1","D29.1","D56.1","D46.1","D12B.1","D67B.1",
                               "D33.2","D29.2","D46.2","D12B.2","D67B.2",
                               "D35.2","D35.3","D56.2","D56.3")) %>%
    ggplot(aes(x=Sample, y=Name))

  ##Plot ggplot2 object, size bubbles is tpm, color is gene
  BubblePlot <- SamplePlot +
    geom_count(aes(size=tpm, color=Pathway)) +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
          panel.background = element_blank(), axis.line = element_line(colour = "black")) +
    scale_size(range = c(1,15))

  return(BubblePlot)

}
SampleRedoxPathBubblePlot <- function(dataframe){
  # return a bubble style plot for a gene abundance per Sample and Sample Grouping
  # Inputs: dataframe = raw data frame of gene abundance data formatted with colnames: Name,
Sample, Group, tpm

  cols <- c('Name', 'Group')
  dataframe[cols] <- lapply(dataframe[cols], factor)

  ##Make $Pathway column

  RedoxList <- c('dsrB','dsrA','aprA','aprB')
  dataframe$Redox <- ifelse(dataframe$Name %in% RedoxList, 'Reduction', 'Oxidation')

  #Order x-axis samples, remove points for samples with zeros
  SamplePlot <- dataframe[which(dataframe$tpm>0),] %>%
    mutate(Sample = fct_relevel(Sample,
                               "D35.1","D33.1","D29.1","D56.1","D46.1","D12B.1","D67B.1",
                               "D33.2","D29.2","D46.2","D12B.2","D67B.2",
                               "D35.2","D35.3","D56.2","D56.3")) %>%
    ggplot(aes(x=Sample, y=Name))

  ##Plot ggplot2 object, size bubbles is tpm, color is gene
  BubblePlot <- SamplePlot +
    geom_count(aes(size=tpm, color=Redox)) +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
          panel.background = element_blank(), axis.line = element_line(colour = "black")) +
    scale_size(range = c(1,15))

  return(BubblePlot)

}
#################################################################################

######################## GroupBubblePlot() ###################################
GroupBubblePlot <- function(dataframe){
  # return a bubble style plot for a gene abundance per Sample and Sample Grouping
  # Inputs: dataframe = raw data frame of gene abundance data formatted with colnames: Name,
Sample, Group, tpm

  cols <- c('Name', 'Group')
  dataframe[cols] <- lapply(dataframe[cols], factor)

  #Make new df of means
  df_means <- aggregate(tpm ~ Group + Name, data=dataframe, FUN=mean)

  df_means[cols] <- lapply(df_means[cols], factor)

  dfMeansPlot <- ggplot(df_means, aes(x = Group, y = Name))

  GroupPlot <- dfMeansPlot +
    geom_count(aes(size = tpm, color=Name)) +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
          panel.background = element_blank(), axis.line = element_line(colour = "black")) +
    scale_size(range = c(1,15))

  return(GroupPlot)

}
GroupRedoxBubblePlot <- function(dataframe){
  # return a bubble style plot for a gene abundance per Sample and Sample Grouping
  # Inputs: dataframe = raw data frame of gene abundance data formatted with colnames: Name,
Sample, Group, tpm
```

```
  cols <- c('Name', 'Group')
  dataframe[cols] <- lapply(dataframe[cols], factor)

  #Make new df of means
  df_means <- aggregate(tpm ~ Group + Name, data=dataframe, FUN=mean)

  RedoxList <- c('dsrB','dsrA','aprA','aprB')
  df_means$Redox <- ifelse(df_means$Name %in% RedoxList, 'Reduction', 'Oxidation')

  df_means[cols] <- lapply(df_means[cols], factor)

  dfMeansPlot <- ggplot(df_means, aes(x = Group, y = Name))

  GroupPlot <- dfMeansPlot +
    geom_count(aes(size = tpm, color=Redox)) +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
          panel.background = element_blank(), axis.line = element_line(colour = "black")) +
    scale_size(range = c(1,15))

  return(GroupPlot)

}
GroupNitPathBubblePlot <- function(dataframe){
  # return a bubble style plot for a gene abundance per Sample and Sample Grouping
  # Inputs: dataframe = raw data frame of gene abundance data formatted with colnames: Name,
Sample, Group, tpm

  cols <- c('Name', 'Group')
  dataframe[cols] <- lapply(dataframe[cols], factor)

  #Make new df of means
  df_means <- aggregate(tpm ~ Group + Name, data=dataframe, FUN=mean)

  S1List <- c('narI','napB','napA')
  S2List1 <- c('narG','narH')
  S3List <- c('norB','norC','norD','norQ')
  NRList <- c('nirB','nirD','nrfA','nrfH')

  for (row in 1:nrow(df_means)) {
    if (df_means$Name[row] %in% S1List) {
      df_means$Pathway[row] <- 'Denitrification/Nitrate Reduction-S1'
    } else if (df_means$Name[row] %in% S2List1) {
      df_means$Pathway[row] <- 'Denitrification/Nitrification-S2'
    } else if  (df_means$Name[row] == 'nirS') {
```

```
   df_means$Pathway[row] <- 'Denitrification-S2'
  } else if  (df_means$Name[row] %in% S3List) {
   df_means$Pathway[row] <- 'Denitrification-S3'
  } else if  (df_means$Name[row] %in% NRList) {
   df_means$Pathway[row] <- 'Nitrate Reduction'
  } else {
   print("Something is wrong")
  }}

 df_means[cols] <- lapply(df_means[cols], factor)

 dfMeansPlot <- df_means[which(df_means$tpm>0),] %>%
  mutate(Name = fct_relevel(Name,
               "napA","napB","narI","narG","narH","nirS","norB","norC","norD",
               "norQ","nirB","nirD","nrfA","nrfH")) %>%
  ggplot(aes(x=Group, y=Name))

 GroupPlot <- dfMeansPlot +
  geom_count(aes(size = tpm, color=Pathway)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
     panel.background = element_blank(), axis.line = element_line(colour = "black")) +
  scale_size(range = c(1,15))

 return(GroupPlot)

}
GroupCPathwayBubblePlot <- function(dataframe){
 # return a bubble style plot for a gene abundance per Sample and Sample Grouping
 # Inputs: dataframe = raw data frame of gene abundance data formatted with colnames: Name,
Sample, Group, tpm

 cols <- c('Name', 'Group')
 dataframe[cols] <- lapply(dataframe[cols], factor)

 #Make new df of means
 df_means <- aggregate(tpm ~ Group + Name, data=dataframe, FUN=mean)

 rbcList <- c('rbcS','rbcL')
 methList <- c('mcrA','mcrB','mcrG')
 aeroCODHList <- c('coxS','coxM','coxL')
 anaeroCODHList <- c('cooS_acsA','cdhE_acsC','cdhD_acsD','cdhC','cdhA','acsB')

 for (row in 1:nrow(df_means)) {
  if (df_means$Name[row] %in% rbcList) {
      df_means$Pathway[row] <- 'RuBiSCO'
    } else if (df_means$Name[row] %in% methList) {
      df_means$Pathway[row] <- 'Methanogenesis'
    } else if  (df_means$Name[row] %in% aeroCODHList) {
      df_means$Pathway[row] <- 'AerobicCODH'
    } else if  (df_means$Name[row] %in% anaeroCODHList) {
      df_means$Pathway[row] <- 'AnaerobicCODH'
    } else {
      print("Something is wrong")
    }}

  df_means[cols] <- lapply(df_means[cols], factor)

  print(df_means$Pathway)

  dfMeansPlot <- df_means[which(df_means$tpm>0),] %>%
    #mutate(Name = fct_relevel(Name,
    #                        "napA","napB","narI","narG","narH","nirS","norB","norC","norD",
    #                        "norQ","nirB","nirD","nrfA","nrfH")) %>%
    ggplot(aes(x=Group, y=Name))

  GroupPlot <- dfMeansPlot +
    geom_count(aes(size = tpm, color=Pathway)) +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
          panel.background = element_blank(), axis.line = element_line(colour = "black")) +
    scale_size(range = c(1,15))

  return(GroupPlot)

}
#############################################################################

######################## Run functions ######################################

SampleBubblePlot(dfSulfurTPM)
GroupBubblePlot(dfSulfurTPM)

SampleBubblePlot(dfDenit1TPM)
GroupBubblePlot(dfDenit1TPM)

SampleBubblePlot(dfDenit2TPM)
GroupBubblePlot(dfDenit2TPM)

SampleBubblePlot(dfCarbonTPM)
```

```
GroupBubblePlot(dfCarbonTPM)

##Finals w/ color here
GroupRedoxBubblePlot(dfSulfurTPM)
GroupCPathwayBubblePlot(dfCarbonTPM)
SampleCPathBubblePlot(dfCarbonTPM)
SampleRedoxPathBubblePlot(dfSulfurTPM)

###############################################################################

################### Change Data in Nitrogen Plots ##############################
#dfNitTPM has $Pathway column added in excel

dfNitTPM_new <- dfNitTPM %>%
  filter(Name != 'norE')

dfNitTPM_noNZ <- dfNitTPM_new %>%
  filter(Name != 'nosZ')

dfNitTPM_noNir <- dfNitTPM_noNZ %>%
  filter(Name != 'nirK')

SampleNitPathBubblePlot(dfNitTPM_noNir)
GroupNitPathBubblePlot(dfNitTPM_noNir)
###############################################################################

#################### MAG "geom_tile()" Plot ####################################

#load data as MAG_Mods

str(MAG_Mods)

## row 505+ are empty
MAG_Mods <- MAG_Mods[1:504,]

#Make ggplot Object

for (row in 1:nrow(MAG_Mods)) {
  if (MAG_Mods$Completeness[row] == 'complete') {
    MAG_Mods$Value[row] <- 1
  } else if (MAG_Mods$Completeness[row] == 'partial') {
    MAG_Mods$Value[row] <- 0.5
  } else if  (MAG_Mods$Completeness[row] == 'none') {
    MAG_Mods$Value[row] <- 0
  } else {
    print("Something is wrong")
  }}

MAG_Mods$Value <- factor(MAG_Mods$Value)
MAG_Mods$Pathway <- factor(MAG_Mods$Pathway)


MAG_Mods <- MAG_Mods %>%
  mutate(Completeness = fct_relevel(Completeness, 'complete', 'partial', 'none'))

MAGPlot <- MAG_Mods %>%
  mutate(Pathway = fct_relevel(Pathway,
                'Complex I',
                'Cyt. c Oxidase',
                'TCA cycle, 1st C Oxidation',
                'TCA cycle, 2nd C Oxidation',
                'RuBisCO',
                'Aerobic CODH',
                'Anaerobic CODH/ACS (acsACD)',
                'Acetyl-CoA Synthetase (acs)',
                'Acetyl-CoA Synthase (acsB)',
                'Acetate Kinase',
                'Phosphate Acetyltransferase',
                'Sox System (soxABCXYZ)',
                'Sulfide:quinone Reductase (SQR)',
                'Nitrate Reductase (napAB or narGHI)',
                'Nitrite Reductase (nirK)',
                'Nitric Oxide Reductase (norBQ)',
                'Nitrous Oxide Reductase (nosZ)',
                'Nitrite Reductase (nirBD or nrfAH)')) %>%
  ggplot(aes(Bin, Pathway, fill=Value))



cols <- c("none" = "grey93", "partial" = "grey63", "complete" = "black")

MAGPlot +
  geom_tile(aes(fill=Completeness), color = 'white') +
  scale_y_discrete(limits=rev) +
  scale_fill_manual(values = cols) +
  ggtitle("Metabolic Pathways and Complexes in BSF MAGs") +
  labs(fill = element_blank()) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.text.x = element_text(angle = 90, vjust = 0.35, size = 12),
        axis.title.x = element_blank(),
        axis.text.y = element_text(size = 12),
        axis.title.y = element_blank(),
        plot.title = element_text(size=21, face="bold", hjust = 0.5))
################################################################################

###################### TPM Heatmap #############################################

RuBisCO <- c('K01601', 'K01602')

anaCO <- c(
  'K00192',
  'K00193',
  'K00195',
  'K15023',
  'K00196',
  'K00197',
  'K00194',
  'K00198',
  'K14138')

aeroCO <- c('K03518', 'K03519', 'K03520')

Acetate <- c('K00925', 'K00625', 'K13788')

Meth <- c('K00399', 'K00401', 'K00402', 'K14083', 'K01895')

Denitrification <- c(
  'K00368',
  'K00376',
  'K04561',
  'K04748',
  'K02567',
  'K02568',
  'K02448',
  'K15864',
  'K00370',
  'K00371',
  'K00374')

N2 <- c('K02586', 'K02588', 'K02591')
```

```
Dsr <- c('K11180','K11181')

SQR <- c('K17218')

Sox <- c(
  'K17222',
  'K17223',
  'K17224',
  'K17225',
  'K17226',
  'K17227')

All <- c(Sox, SQR, Dsr, N2, Denitrification, Meth, Acetate, aeroCO, anaCO, RuBisCO)

overviewTPM <- samples_INT_tpm %>%
  filter(KEGG_ID %in%  All)

overviewTPM <- overviewTPM[,1:18]

for (row in 1:nrow(overviewTPM)) {
  if (overviewTPM$KEGG_ID[row] %in% Sox) {
    overviewTPM$Definition[row] <- 'SOX System'
  } else if (overviewTPM$KEGG_ID[row] %in% SQR) {
    overviewTPM$Definition[row] <- 'Sulfide Oxidation'
  } else if  (overviewTPM$KEGG_ID[row] %in% Dsr) {
    overviewTPM$Definition[row] <- 'Sulfate Reduction'
  } else if  (overviewTPM$KEGG_ID[row] %in% N2) {
    overviewTPM$Definition[row] <- 'N2 Fixation'
  } else if  (overviewTPM$KEGG_ID[row] %in% Denitrification) {
    overviewTPM$Definition[row] <- 'Denitrification'
  } else if  (overviewTPM$KEGG_ID[row] %in% Meth) {
    overviewTPM$Definition[row] <- 'Methanogenesis'
  } else if  (overviewTPM$KEGG_ID[row] %in% Acetate) {
    overviewTPM$Definition[row] <- 'Acetate Formation'
  } else if  (overviewTPM$KEGG_ID[row] %in% aeroCO) {
    overviewTPM$Definition[row] <- 'Aerobic CODH'
  } else if  (overviewTPM$KEGG_ID[row] %in% anaCO) {
    overviewTPM$Definition[row] <- 'Anaerobic CODH'
  } else if  (overviewTPM$KEGG_ID[row] %in% RuBisCO) {
    overviewTPM$Definition[row] <- 'RuBisCO'
  } else {
    print("Something is wrong")
  }}
```

```
overviewTPM <- overviewTPM %>%
  mutate(Definition = fct_relevel(Definition,
                              'Sulfide Oxidation',
                              'Sulfate Reduction',
                              'SOX System',
                              'Aerobic CODH',
                              'Anaerobic CODH',
                              'Acetate Formation',
                              'RuBisCO',
                              'Denitrification',
                              'N2 Fixation',
                              'Methanogenesis'))


overviewTPM[9,]$Name <- 'narG/nxrA'
overviewTPM[10,]$Name <- 'narH/nxrB'
overviewTPM[11,]$Name <- 'narI'
overviewTPM[34,]$Name <- 'pta 1'
overviewTPM[6,]$Name <- 'acsC'
overviewTPM[7,]$Name <- 'acsA'
overviewTPM[3,]$Name <- 'acsD'
overviewTPM[20,]$Name <- 'acs'
rownames(overviewTPM)
rownames(overviewTPM) <- c(overviewTPM$Name)

overviewTPM$Name

overviewMatrix <- as.matrix(overviewTPM[,2:17])

overviewScaledMatrix <- scale(overviewMatrix)

#### sample_groups is two cols: Sample and Group
groups.df <- as.data.frame(sample_groups[,2], row.names = sample_groups$Sample)

groups.df

groups.t.df <- t(groups.df)

groups.t.df[1,1:16]

Heatmap(overviewScaledMatrix,
        row_split = overviewTPM$Definition,
        column_split = groups.t.df[1,1:16],
        row_names_gp = gpar(fontsize = 11))

######################################################################################
```