Integrating molecular biology and geochemical tools to explore sulfur and methane metabolism in the Santa Monica Basin, CA

Kat Dawson

Feb 2017

Load and/or install needed packages

```
library(pacman)
pacman::p_load(reshape2, ggplot2, vegan, gtools, stats, superheat, RColorBrewer, Heatplus, gplots, corr
```

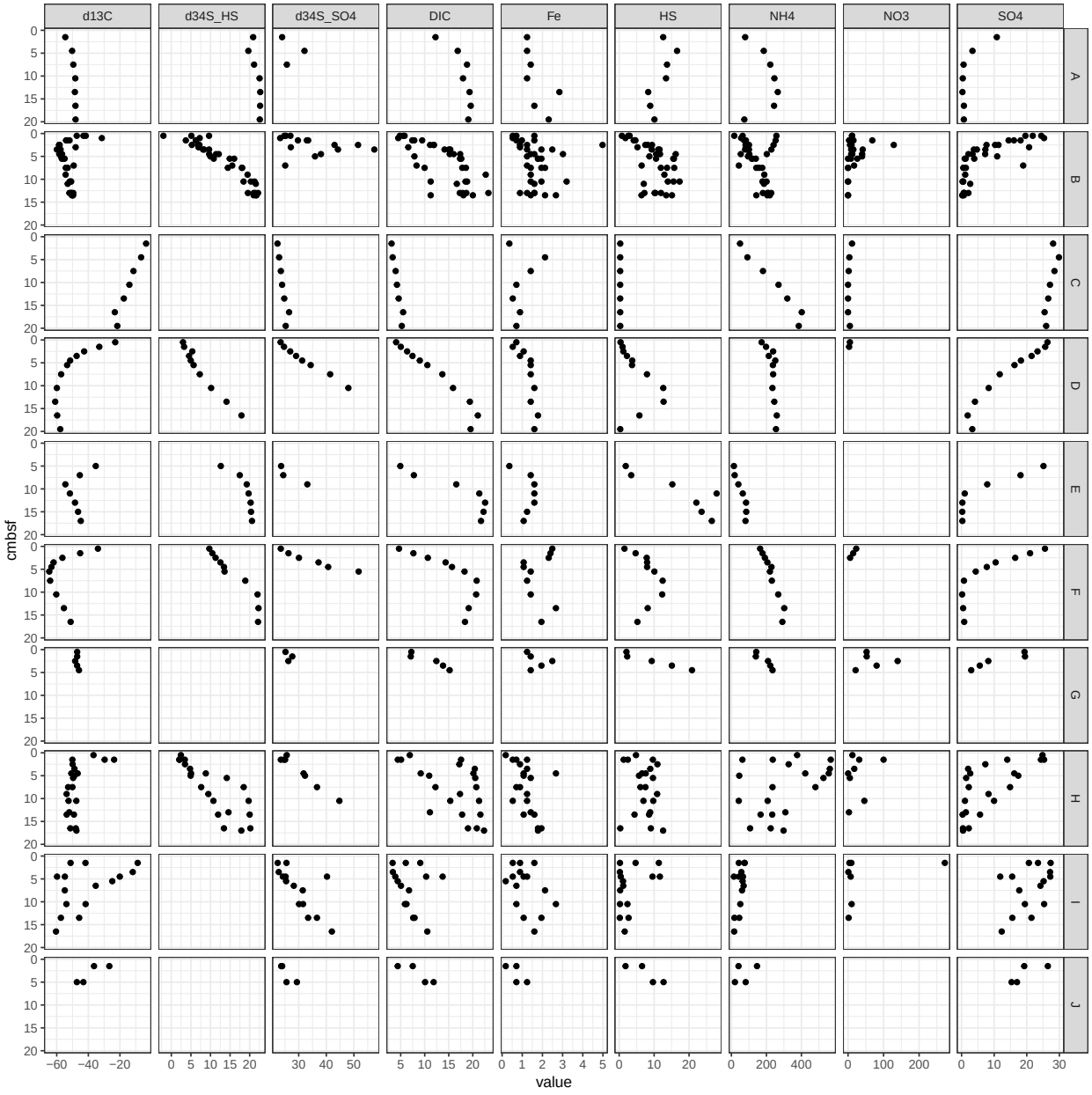
Geochemistry overview:

melt geochem data for plotting

```
#setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
#geochem=read.table("SMB_geochem.txt",header=TRUE)
#geochem2=melt(geochem,id.vars=c("Sample","Site","Type2","Type","cmbsf"),na.rm=FALSE)
#write.table(geochem2,file="SMB_geochem_melt.txt",sep="\t")
```

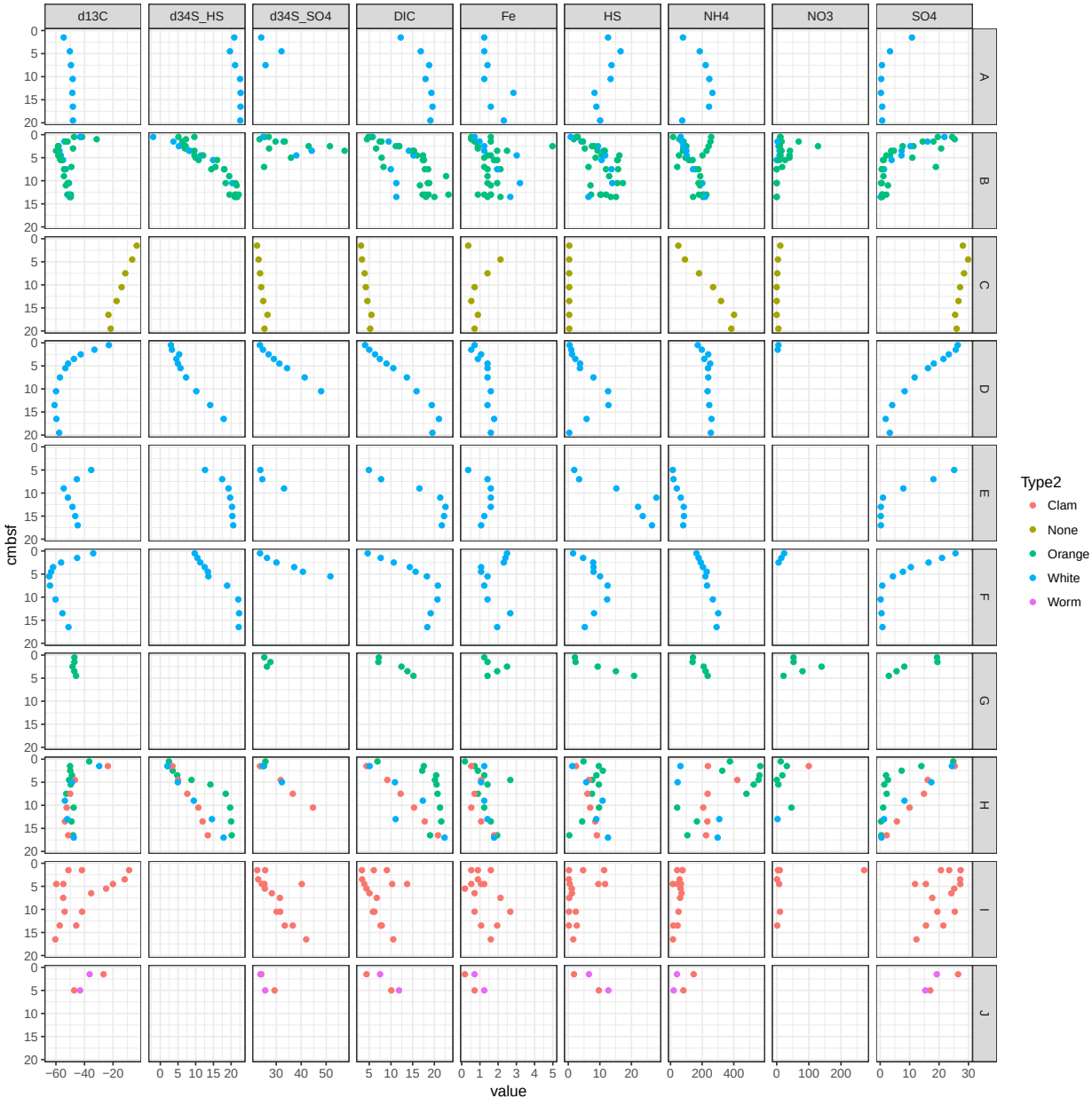
Santa Monica Basin Geochemistry by sampling site

```
setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
geochem2=read.table("SMB_geochem2.txt",header=TRUE)
ggplot(data=na.omit(geochem2),aes(x=value,y=cmbsf))+geom_point()+facet_grid(Site~variable,scales="free_
```



Santa Monica Basin Geochemistry color coded by environment type

```
setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
geochem2=read.table("SMB_geochem2.txt",header=TRUE)
ggplot(data=na.omit(geochem2),aes(x=value,y=cmbsf,color=Type2))+geom_point()+facet_grid(Site~variable,s
```

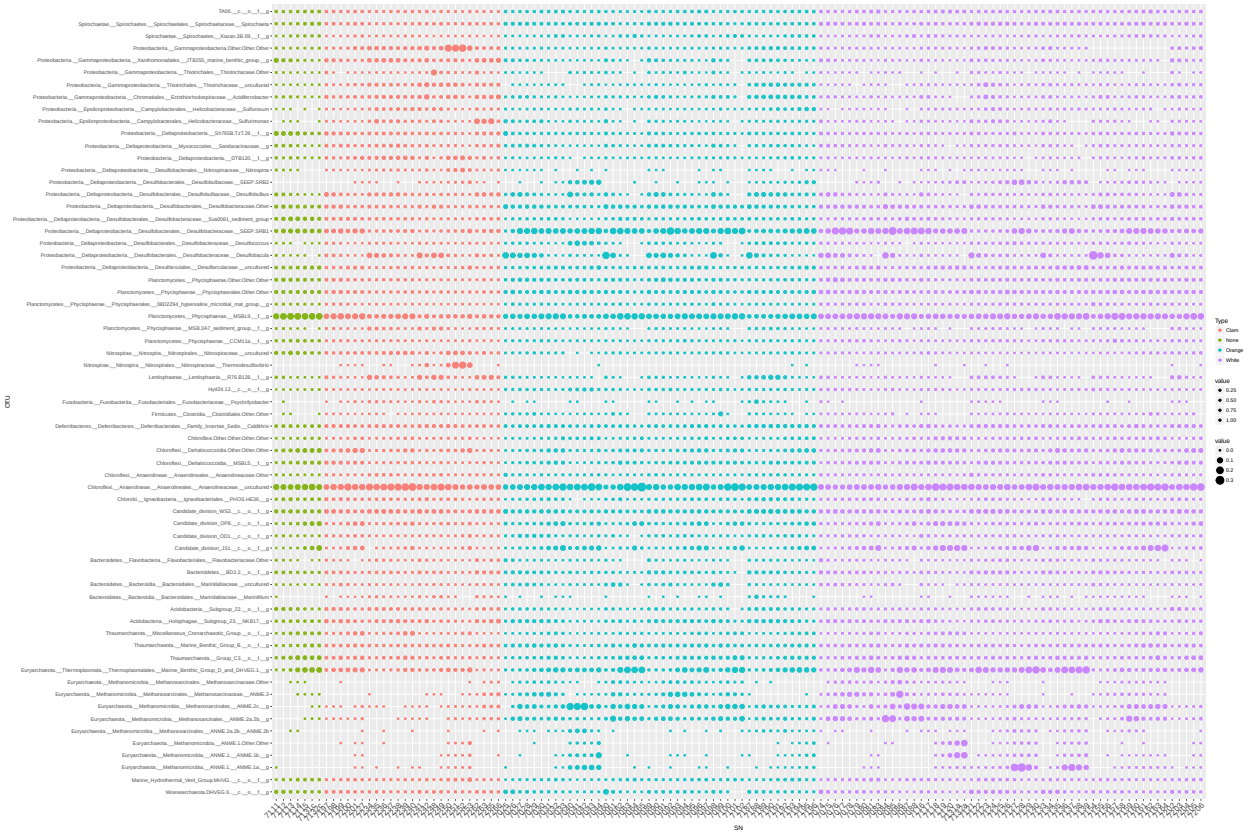


OTU genera overview:

Santa Monica Basin, OTU genera with 2% abundance cutoff

```
setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
dataset1=read.table("SM2013_L6_w_meta2.txt",header=TRUE)
#Set cut off for % abundance, currently set to 2% abundance in at least one sample
cutoff=0.02
dataset1_meta=dataset1[,1:5]
row.names(dataset1_meta)=dataset1[,1]
#Set order for samples (Be aware: ggplot is very literal. A1 is followed by A10 not A2. So, use: A001,.
dataset1_meta$SN=factor(dataset1_meta$SN,levels=dataset1_meta$SN[order(dataset1_meta$Order)])
```

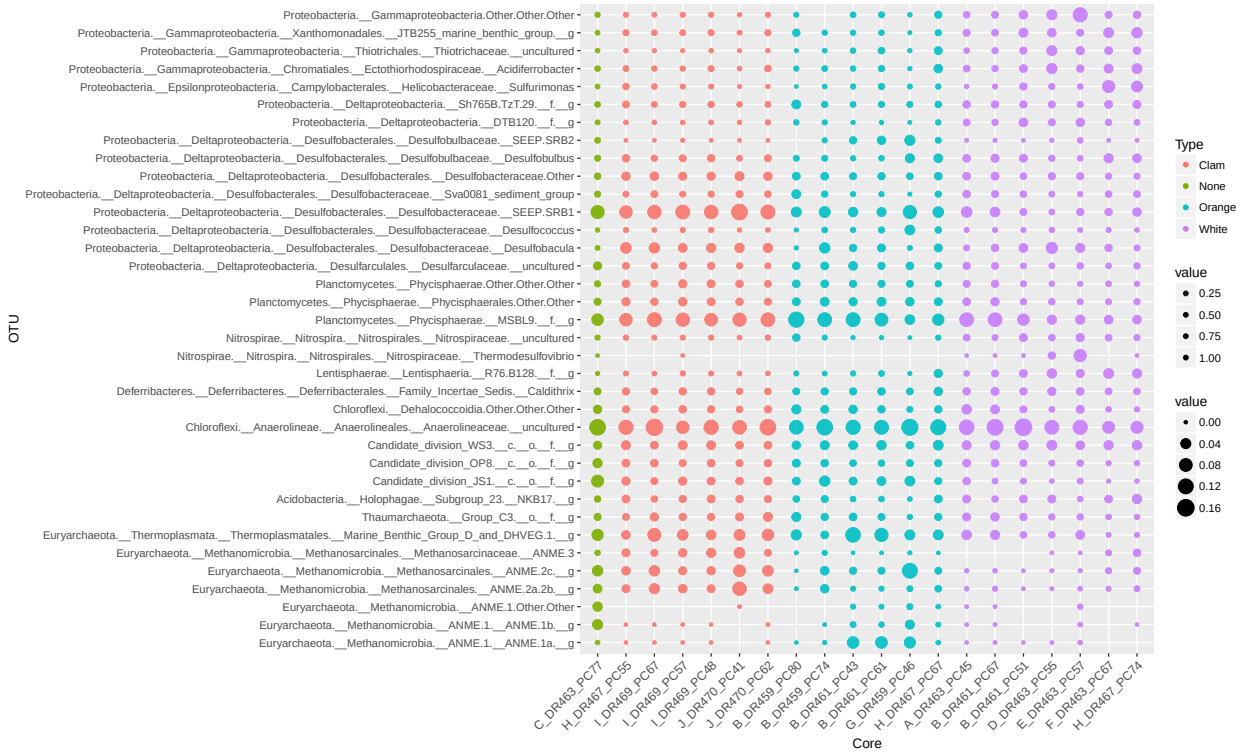
```
#reduce OTU file by minimum OTU abundance threshold
row.names(dataset1)=dataset1[,1]
tagdata.1=dataset1[,-(1:5)]
data_matrix1=tagdata.1/100
maxab=apply(data_matrix1,2,max)
n1=names(which(maxab>cutoff))
data=data_matrix1[,which(names(data_matrix1) %in% n1)]
#write.table(data,file="tagdata_more_than_x_percent.txt",sep="\t")
data1=as.matrix(data)
data1_meta=merge(data1,dataset1_meta,by="row.names")
#melt dataframe
data_new <- melt(data1_meta)
data_new.1=data_new[,-1]
names(data_new.1)=c("SN","Site","Order","Core","Type","OTU","value")
#Plot data. With na.value=0, no circle is made if the OTU is absent in that sample.
p1=ggplot(data_new.1, aes(x=SN, y=OTU)) +geom_point(aes(size=value,colour=Type,alpha=value))+scale_alpha
p1
```



Santa Monica Basin, OTU genera (2% cutoff) core average overview

```
setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
dataset1=read.table("SM2013_core_avg_w_meta2.txt",header=TRUE)
#Set cut off for % abundance, currently set to 2% abundance in at least one sample
cutoff=0.02
dataset1_meta=dataset1[,1:4]
```

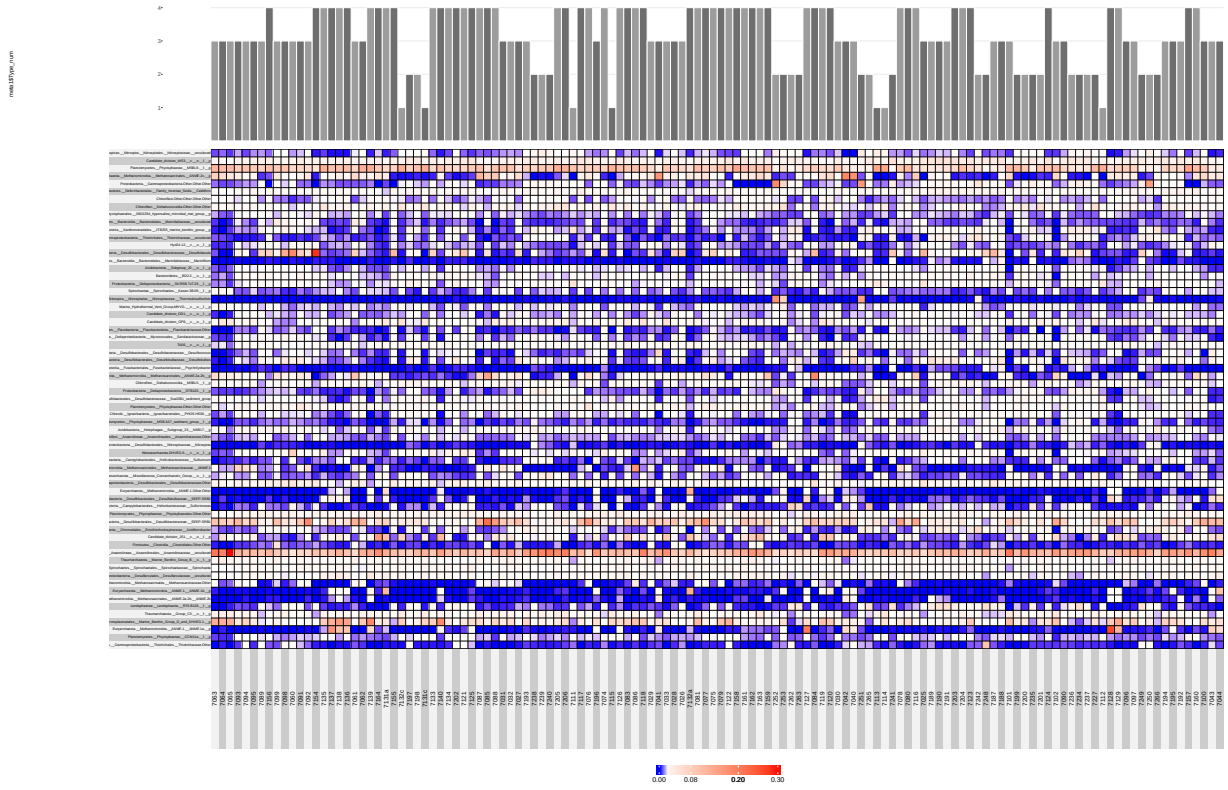
```
row.names(dataset1_meta)=dataset1[,1]
#Set order for samples (Be aware: ggplot is very literal. A1 is followed by A10 not A2. So, use: A001,.
dataset1_meta$Core=factor(dataset1_meta$Core,levels=dataset1_meta$Core[order(dataset1_meta$Order)])
#reduce OTU file by minimum OTU abundance threshold
row.names(dataset1)=dataset1[,1]
tagdata.1=dataset1[,-(1:4)]
data_matrix1=tagdata.1/100
maxab=apply(data_matrix1,2,max)
n1=names(which(maxab>cutoff))
data=data_matrix1[,which(names(data_matrix1) %in% n1)]
write.table(data,file="tagdata_more_than_x_percent.txt",sep="\t")
data1=as.matrix(data)
data1_meta=merge(data1,dataset1_meta,by="row.names")
#melt dataframe
data_new <- melt(data1_meta)
data_new.1=data_new[,-1]
names(data_new.1)=c("Core","Site","Order","Type","OTU","value")
#plot data
p2=ggplot(data_new.1, aes(x=Core, y=OTU))+geom_point(aes(size=value,colour=Type,alpha=value))+scale_alpl
p2
```



Santa Monica Basin, OTU genera (2% cutoff) heatmap

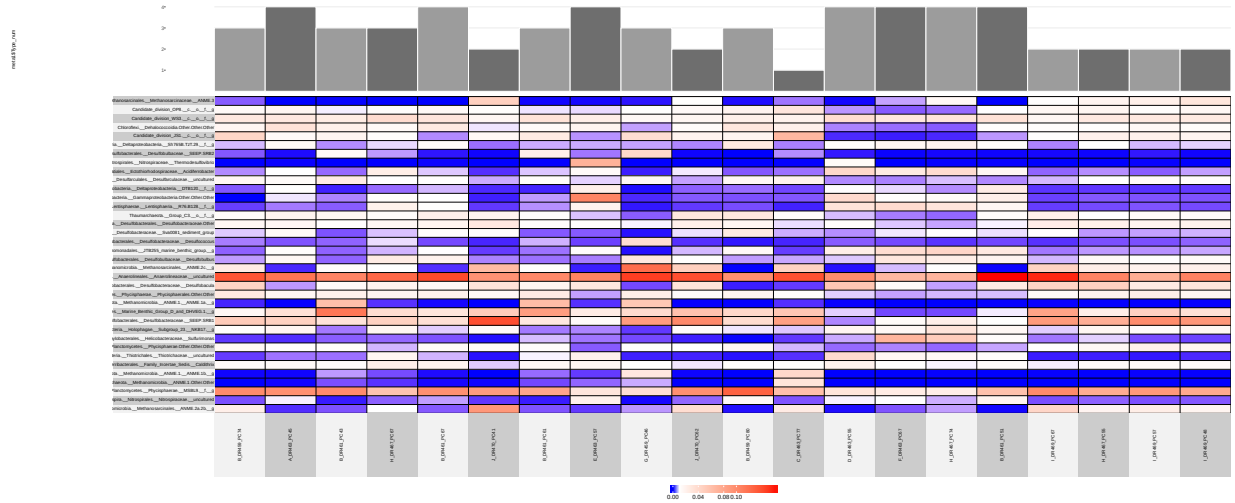
```
#For more information: https://rlbarter.github.io/superheat/index.html
setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
dataset1=read.table("SM2013_L6_w_meta2.txt",header=TRUE)
meta1=read.table("SMB2013_L6all_meta.txt",header=TRUE)
# This code assumes your data is a matrix oriented with samples in descending rows and OTUs in columns
```

```r
# If that's not the case, use the commented out code below to transpose the matrix
# dataset1=t(dataset1)
#Set cut off for % abundance, currently set to 2% abundance in at least one sample
cutoff=0.02
#If you want to set the order for samples (Be aware: ggplot is very literal. A1 is followed by A10 not A
#dataset1$SN=factor(dataset1$SN,levels=dataset1$SN[order(dataset1$Order)])
#reduce OTU file by minimum OTU abundance threshold
row.names(dataset1)=dataset1[,1]
tagdata.1=dataset1[,-(1:6)]
data_matrix1=tagdata.1/100
maxab=apply(data_matrix1,2,max)
n1=names(which(maxab>cutoff))
data=data_matrix1[,which(names(data_matrix1) %in% n1)]
data_t=t(data)
#Cluster data
data.dist=vegdist(data_t,method="euclidean")
row.clus=hclust(data.dist,"aver")
data.dist.g=vegdist(t(data_t),method="euclidean")
col.clus=hclust(data.dist.g,"aver")
#plot heatmap
#To add if you want gridlines: grid.hline.col="white",grid.hline.size=1,grid.vline.col="white",grid.vli
superheat(data_t,left.label.text.size=2,left.label.text.alignment="right", force.bottom.label=TRUE, bott
```



Santa Monica Basin, OTU genera core average (2% cutoff) heatmap

```
#For more information: https://rlbarter.github.io/superheat/index.html
setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
dataset1=read.table("SM2013_core_avg_w_meta2.txt",header=TRUE)
meta1=read.table("SMB2013_L6avg_meta.txt",header=TRUE)
# This code assumes your data is a matrix oriented with samples in descending rows and OTUs in columns
# If that's not the case, use the commented out code below to transpose the matrix
# dataset1=t(dataset1)
#Set cut off for % abundance, currently set to 2% abundance in at least one sample
cutoff=0.02
#reduce OTU file by minimum OTU abundance threshold
row.names(dataset1)=dataset1[,1]
tagdata.1=dataset1[,-(1:6)]
data_matrix1=tagdata.1/100
maxab=apply(data_matrix1,2,max)
n1=names(which(maxab>cutoff))
data=data_matrix1[,which(names(data_matrix1) %in% n1)]
data_t=t(data)
#Cluster data
data.dist=vegdist(data_t,method="euclidean")
row.clus=hclust(data.dist,"aver")
data.dist.g=vegdist(t(data_t),method="euclidean")
col.clus=hclust(data.dist.g,"aver")
#plot heatmap
#To add if you want gridlines: grid.hline.col="white",grid.hline.size=1,grid.vline.col="white",grid.vli
#yt=t(dataset1$Depth), yt.axis.name="Depth", yt.plot.type="scatter")
superheat(data_t,left.label.text.size=3,left.label.text.alignment="right", force.bottom.label=TRUE, bot
```

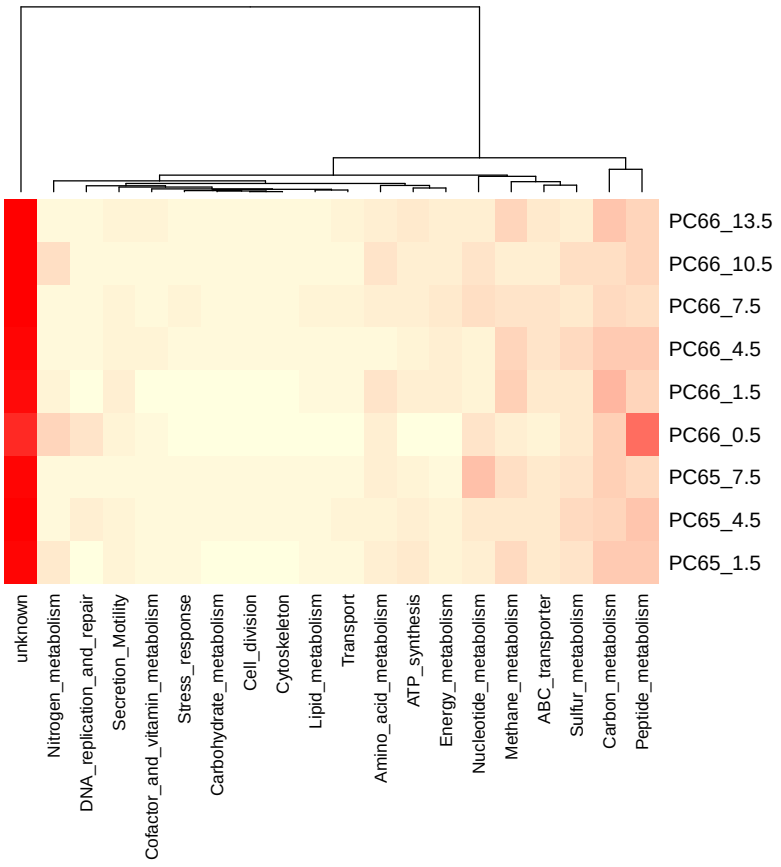


Proteomics overview:

Comparison of Consensus, De Novo, and Quantitative datasets

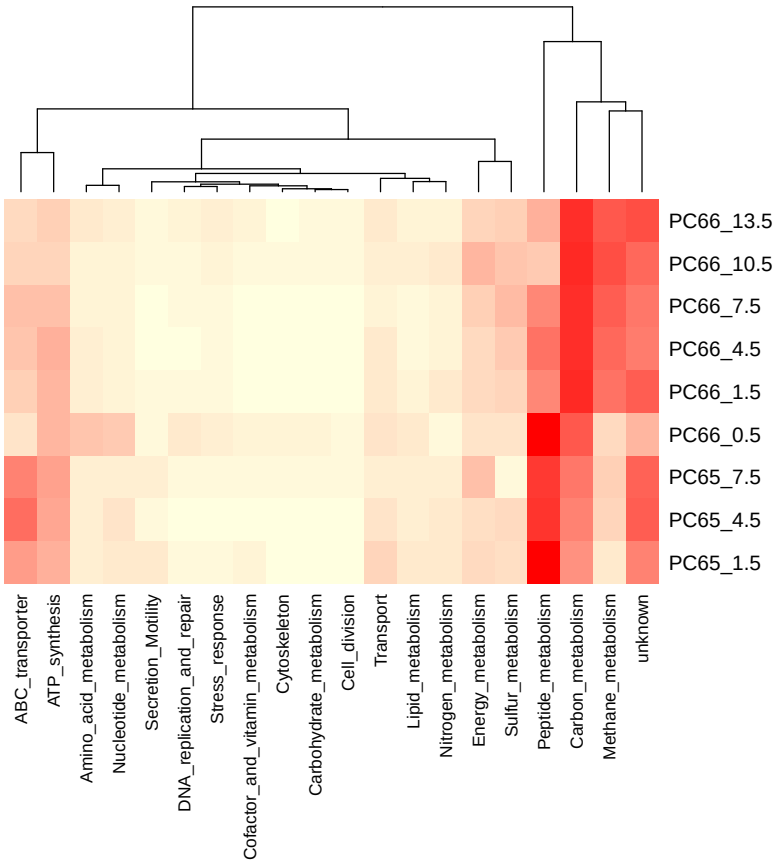
Heatmap of consensus proteome gene ontology

```
setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
data.consensus=read.table("SM2013_proteomeconsensus.txt",header=TRUE, row.names=1)

#Cluster analysis with clustering of variables by sample
# Other distance metrics include "euclidean", "manhattan","gower", "canberra","kulczynski,"morsita,"hor

data.dist=vegdist(data.consensus,method="euclidean")
row.clus=hclust(data.dist,"aver")
data.dist.g=vegdist(t(data.consensus),method="euclidean")
col.clus=hclust(data.dist.g,"aver")

# Generate a heatmap and dendrogram, with a yellow to red color scale change these colors as desired; t

scaleyellowred=colorRampPalette(c("lightyellow","red"),space="rgb")(50)

#Simple heatmap with dendrograms

heatmap(as.matrix(data.consensus),Colv=as.dendrogram(col.clus),Rowv=NA,col=scaleyellowred,margins=c(20,
```



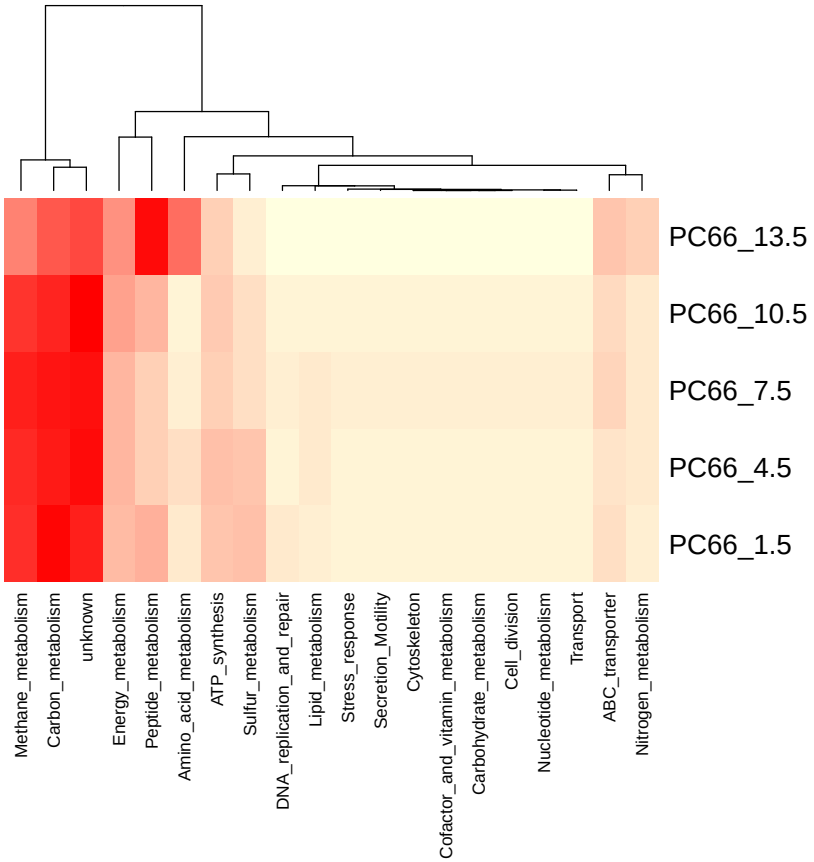
Heatmap of de novo proteome gene ontology

```
setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
data.denovo=read.table("SM2013_proteomedenovo.txt",header=TRUE, row.names=1)

#Cluster analysis with clustering of variables by sample
# Other distance metrics include "euclidean", "manhattan","gower", "canberra","kulczynski,"morsita,"hor

data.dist=vegdist(data.denovo,method="euclidean")
row.clus=hclust(data.dist,"aver")
data.dist.g=vegdist(t(data.denovo),method="euclidean")
col.clus=hclust(data.dist.g,"aver")

# Generate a heatmap and dendrogram, with a yellow to red color scale change these colors as desired; t

scaleyellowred=colorRampPalette(c("lightyellow","red"),space="rgb")(50)

#Simple heatmap with dendrograms

heatmap(as.matrix(data.denovo),Colv=as.dendrogram(col.clus),Rowv=NA,col=scaleyellowred,margins=c(20,7))
```



Heatmap of quantitative proteome gene ontology

```
setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
data.quant=read.table("SM2013_proteomequant.txt",header=TRUE, row.names=1)
#Cluster analysis with clustering of variables by sample
# Other distance metrics include "euclidean", "manhattan","gower", "canberra","kulczynski,"morsita,"hor
data.dist=vegdist(data.quant,method="euclidean")
row.clus=hclust(data.dist,"aver")
data.dist.g=vegdist(t(data.quant),method="euclidean")
col.clus=hclust(data.dist.g,"aver")
# Generate a heatmap and dendrogram, with a yellow to red color scale change these colors as desired; t
scaleyellowred=colorRampPalette(c("lightyellow","red"),space="rgb")(50)
#Simple heatmap with dendrograms
heatmap(as.matrix(data.quant),Colv=as.dendrogram(col.clus),Rowv=NA,col=scaleyellowred,margins=c(20,7))
```

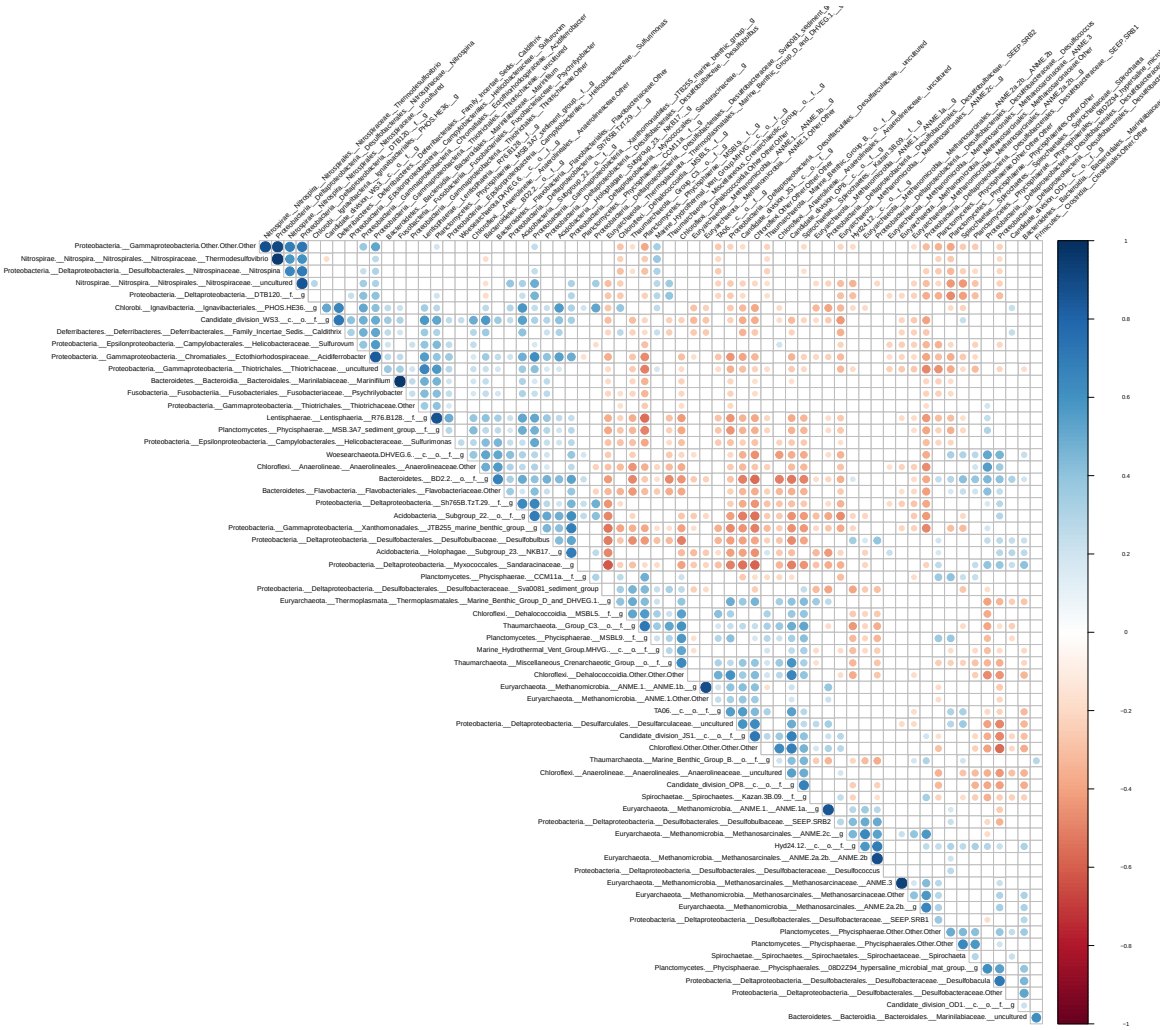


Corelation Analysis

OTU (L6, 2% cutoff) Pearson pairwise correlation in prep for network analysis

```
setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
dataset1=read.table("SM2013_L6_w_meta2.txt",header=TRUE)
#Set cut off for % abundance, currently set to 2% abundance in at least one sample
```

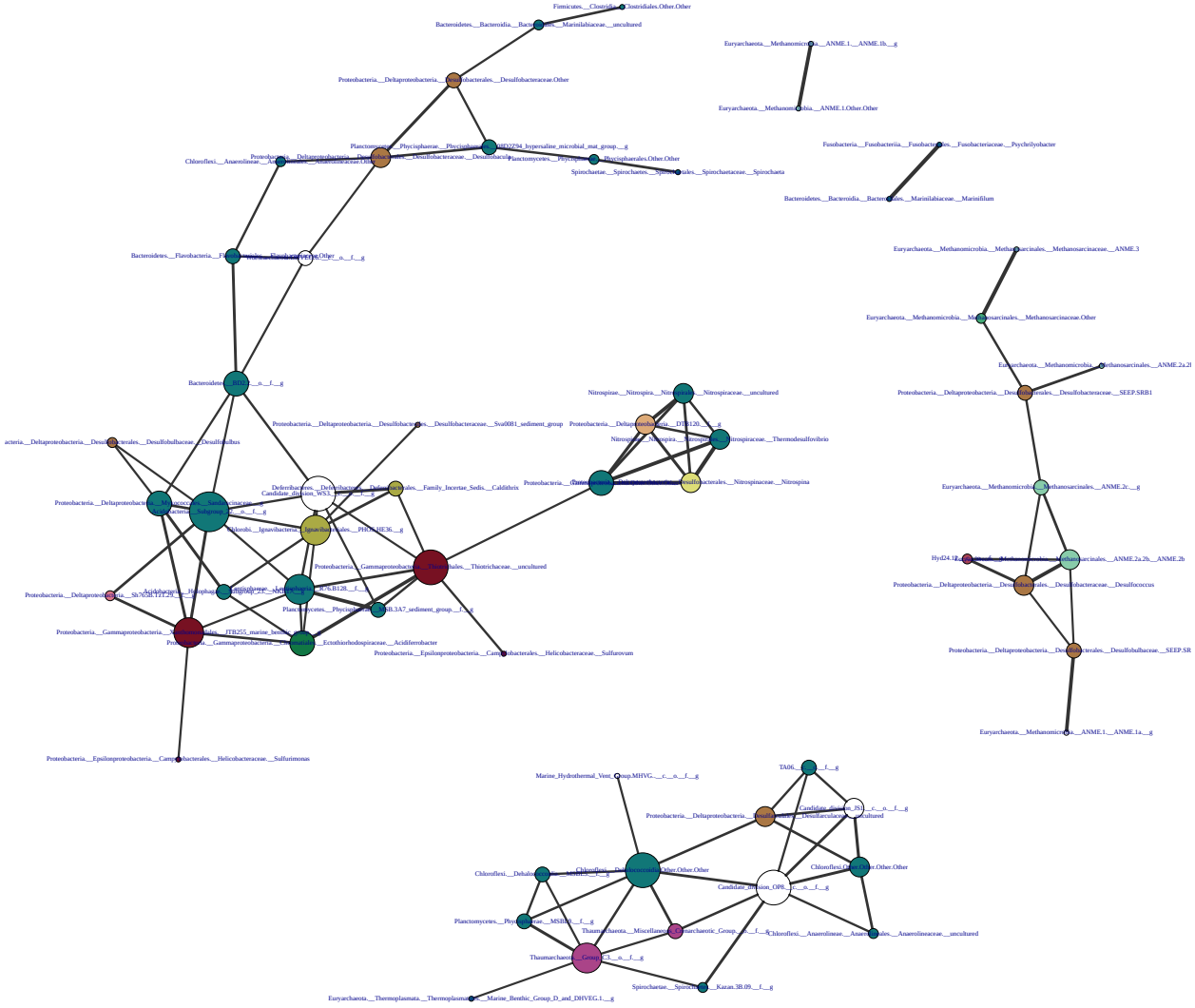
```
cutoff=0.02
dataset1_meta=dataset1[,1:6]
row.names(dataset1_meta)=dataset1[,1]
row.names(dataset1)=dataset1[,1]
tagdata.1=dataset1[,-(1:6)]
data_matrix1=tagdata.1/100
maxab=apply(data_matrix1,2,max)
#head(maxab)
n1=names(which(maxab>cutoff))
data=data_matrix1[,which(names(data_matrix1) %in% n1)]
data1=as.matrix(data)
#write.table(data,file=paste(Sys.time(),"tagdata_more_than_x_percent.txt",sep="\t")
#correlation with pvalue pearson or spearman
data.pearsoncor=rcorr(data1,type="pearson")
pearson_pval=data.pearsoncor$P
#write.table(data.pearsoncor$r,file="SM2013_pearsoncor.txt",sep="\t")
#write.table(pearson_pval,file="SM2013_pearson_pval.txt",sep="\t")
corrplot(data.pearsoncor$r, method="circle", type="upper", order="hclust", hclust.method="ward.D", diag=
```



Converting correlation matrix into a network diagram

```
# The above correlation analysis needs to be run first
# Reshape matrices for cytoscape input and rbind p-values with corr coefficients
# Final format will be:
#   OTU1 OTU2  Pearson Pearson_pval  Spearman  Spearman_pval
#   A1    A2    0.3     0.001          0.31      0.0015
setwd("C:/Users/Kat/Desktop/Caltech/Manuscripts/Santa Monica basin/R/R markdown file")
pearson_cor_tri=upper.tri(data.pearsoncor$r,diag=F)
data.pearsoncor_copy=data.pearsoncor$r
data.pearsoncor_copy[!pearson_cor_tri]=NA
data.pearsoncor_melted=na.omit(melt(data.pearsoncor_copy,value.name="correlationCoef"))
```

```
colnames(data.pearsoncor_melted)=c("OTU1","OTU2","Pearson")
# write.csv(data.pearsoncor_melted,"data.pearsoncor_3col.txt")
pearson_pval_tri=upper.tri(pearson_pval,diag=F)
data.pearsoncor_copy2=pearson_pval
data.pearsoncor_copy2[!pearson_pval_tri]=NA
pearson_pval_melted=na.omit(melt(data.pearsoncor_copy2,value.name="pvalue"))
colnames(pearson_pval_melted)=c("OTU1","OTU2","Pearson_pval")
pearson_all=merge(data.pearsoncor_melted,pearson_pval_melted)
# write.csv(pearson_all,"pearson_all_pval.txt")
# subset data to include only pairings with p-val < 0.01
pearson_all_sig=subset(pearson_all,pearson_all$Pearson_pval<0.01)
# Output an excel file that can be used in cytoscape
# Write.xlsx(pearson_all_sig,file="SM2013_corr_sig_pval_for_cytoscape.xlsx")
# Now for converting the correlation file into an edge list and making a network
# For more info see: http://www.kateto.net/wp-content/uploads/2015/06/Polnet%202015%20Network%20Viz%20T
# Pick a threshold for what pearson correlation level will be plotted
pearson_highpos=subset(pearson_all_sig[,1:3],pearson_all_sig$Pearson>0.5)
#pearson_highpos1=pearson_highpos[,c(2,1,3)]
#names(pearson_highpos1)=c("OTU1","OTU2","Pearson")
#pearson_highpos2=rbind(pearson_highpos,pearson_highpos1)
edges_highpos1=as.matrix(pearson_highpos[,1:3])
net=graph_from_data_frame(edges_highpos1,directed=FALSE)
# To color nodes by metabolism type, upload an additional file with that information
# OTU1  Metabolism
# Deltaproteobacteria Sulfate_Reduction
nodes=read.table("SM2013_node_character.txt",header=TRUE)
# Parse nodes list to include only those meeting the Pearson cutoff
# !!! This new table with node metabolic characteristics needs to have all of the ';', '(', ')', and '-
nodes_highpos=c(as.character(pearson_highpos$OTU1),as.character(pearson_highpos$OTU2))
nodes_highpos=as.data.frame(nodes_highpos)
nodes_highpos1=nodes_highpos[!duplicated(nodes_highpos),]
nodes_highpos1=as.data.frame(nodes_highpos[!duplicated(nodes_highpos),])
names(nodes_highpos1)=c("OTU1")
#write.table(nodes_highpos1,file="nodes_highpos.txt",sep="\t")
nodes2=nodes[order(nodes$OTU1),]
nodes_highpos2=join(nodes_highpos1,nodes2,by="OTU1")
V(net)$size=degree(net)
tol21rainbow=c("#771155", "#AA4488", "#CC99BB", "#114477", "#4477AA", "#77AADD", "#117777", "#44AAAA",
colrs=tol21rainbow
V(net)$color=colrs[nodes_highpos2$Metabolism]
E(net)$width=10*as.numeric(edges_highpos1[,3])
plot.igraph(net, layout=layout.fruchterman.reingold, edge.color="gray20")
```


Firmicutes__Clostridia__Clostridiales.Other.Other
Bacteroidetes__Bacteroidia__Bacteroidales__Marinilabiaceae__uncultured
Euryarchaeota__Methanomicrobia__ANME.1__ANME.1b__g
Proteobacteria__Deltaproteobacteria__Desulfobacterales__Desulfobacteraceae.Other
Euryarchaeota__Methanomicrobia__ANME.1.Other.Other
Planctomycetes__Phycisphaerae__Phycisphaerales__MSBL9__f__g
Planctomycetes__Phycisphaerae__Phycisphaerales__SED294_hypersaline_microbial_mat_group__g
Fusobacteria__Fusobacteriia__Fusobacteriales__Fusobacteriaceae__Psychrilyobacter
Chloroflexi__Anaerolineae__Anaerolineales__Anaerolineaceae.Other
Planctomycetes__Phycisphaerae__Phycisphaerales.Other.Other
Spirochaetae__Spirochaetes__Spirochaetales__Spirochaetaceae__Spirochaeta
Bacteroidetes__Bacteroidia__Bacteroidales__Marinilabiaceae__Marinifilum
Bacteroidetes__Flavobacteriia__Flavobacteriales__Flavobacteriaceae.Other__o__f__g
Euryarchaeota__Methanomicrobia__Methanosarcinales__Methanosarcinaceae__ANME.3
Euryarchaeota__Methanomicrobia__Methanosarcinales__Methanosarcinaceae.Other
Euryarchaeota__Methanomicrobia__Methanosarcinales__ANME.2a.2b
Bacteroidetes__o__f__g
Nitrospinae__Nitrospira__Nitrospirales__Nitrospiraceae__uncultured
Proteobacteria__Deltaproteobacteria__Desulfobacterales__Desulfobacteraceae__Sva0081_sediment_group
Proteobacteria__Deltaproteobacteria__Desulfobacterales__SEEP.SRB1
acteria__Deltaproteobacteria__Desulfobacterales__Desulfobulbaceae__Desulfobulbus
Nitrospinae__Nitrospira__Nitrospirales__Nitrospiraceae__Thermodesulfovibrio
Euryarchaeota__Methanomicrobia__Methanosarcinales__ANME.2c__g
Proteobacteria__Deltaproteobacteria__Desulfobacterales__Desulfobacteraceae__Desulfococcus
Proteobacteria__Deltaproteobacteria__Desulfobulbaceae__SEEP.SR
Euryarchaeota__Methanomicrobia__ANME.1__ANME.1a__g
Proteobacteria__Epsilonproteobacteria__Campylobacterales__Helicobacteraceae__Sulfurimonas
Marine_Hydrothermal_Vent_Group.MHVG__c__o__f__g
TA06__f__g
Proteobacteria__Deltaproteobacteria__Desulfobacterales__Desulfobacteraceae__uncultured
Chloroflexi__Dehalococcoidia__SAR202_clade__f__g
Chloroflexi__Dehalococcoidia__Dehalococcoidia.Other.Other
Candidate_division_OP8__c__o__f__g
Planctomycetes__Phycisphaerae__MSBL9__f__g
Thaumarchaeota__Miscellaneous_Crenarchaeotic_Group__c__f__g
Chloroflexi__Anaerolineae__Anaerolineales__Anaerolineaceae__uncultured
Thaumarchaeota__Group_1__o__f__g
Spirochaetae__Spirochaetes__Kazan.3B.09__f__g
Euryarchaeota__Thermoplasmata__Thermoplasmatales__Marine_Benthic_Group_D_and_DHVEG.1__g