Analysis of soil microbial traits and carbon cycling

Overview and goals

This R script give an outline some of the analyses used for the paper “A genomic perspective on stoichiometric regulation of soil carbon cycling” by Hartman et al. 2017 ISME J. in press. Soil carbon (C) cycling, nutrients and microbes were compared across a series of rice fields spanning much of the global range in soil C.

This analysis show that soil C cycling is linked to the availability of carbon (C), nitrogen (N), and phosphorus (P). This ecosystem scale element coupling arises from underlying tradeoffs in C, N, and P use in microbial genomes.

I use three data sets with observations per soil: chemistry, genes, and microbes. The fourth data set is key for this analysis: genes by microbes. This allows assessment of how shifts in ecosystem function reflects different combinations of genes in microbes related to their evolutionary history.

Outline:

1) Soil chemistry and C cycling

2) Soil genes and C cycling

3) Microbial traits linked to C cycling

Import required packages:

```
library(Hmisc)
library(RColorBrewer)
library(gplots)
library(ggplot2)
library(gridExtra)
```

Import datasets

```
soil.chem <- read.table("Soils.txt", header=TRUE, sep="\t")
soil.genes <- read.table("Soil_genes.txt", header=TRUE, sep="\t")
CNP.gene.hier <- read.table("CNP_genes.txt", header=TRUE, sep="\t")
```

2) Soil genes and C cycling

Which genes are most closely related to soil C turnover?

a) Get C, N, P cycling genes and determine which are correlated

```
# Select only genes used for C, N, and P cycling
CNP.genes <-data.frame("Gene"=CNP.gene.hier$Gene) # Get list of genes
CNP.gene.abund <- merge(CNP.genes, soil.genes, by="Gene", all.x=TRUE) # merge to get onl
y CNP genes
  row.names(CNP.gene.abund)<-CNP.gene.abund$Gene # set rownames before transpose

# Append soil C turnover data to gene abundance to test which genes are correlated
C.turn <- soil.chem$C_turnover  # get C turnover / soil
CNP.gene.C.turn <- cbind(C.turn, t(CNP.gene.abund[,-1])) # combine w./ gene abund, drop
 gene column
```

Make correlation matrix and retain correlated genes

```
# Get correlation square matrix for r and P values, only use first row
CNP.gene.corr.M <- rcorr(CNP.gene.C.turn)   # test correlations
CNP.gene.corr.m <- data.frame("r"=CNP.gene.corr.M$r[-1,1], "P"=CNP.gene.corr.M$P[-1,1])
# combine r and P
CNP.gene.corr.m$P.fdr <- p.adjust(CNP.gene.corr.m$P, method="BH") # correct for false di
scov. rate

# Filter to make list of genes correlated to C turnover (r > 0.5, P.adj < 0.05)
CNP.gene.corr.Pfilt <- subset(CNP.gene.corr.m, P.fdr < 0.05)
CNP.gene.corr.P.r.filt <- subset(CNP.gene.corr.Pfilt, abs(r) > 0.5)
CNP.gene.corr.list <- data.frame("Gene"= row.names(CNP.gene.corr.P.r.filt)) # extract co
rr. genes
```

b) Organize these genes by function and get their abundance in soil

```
# Gather class/category info for each correlated gene
Gene.corr.hier<-merge(CNP.gene.corr.list,CNP.gene.hier, by="Gene", all.x=TRUE)

# Add soil abundance data, order on gene index
Gene.corr.hier.abund <-merge(Gene.corr.hier,CNP.gene.abund, by="Gene", all.x=TRUE) # mer
ge
  Gene.corr.hier.abund<-Gene.corr.hier.abund[order(Gene.corr.hier.abund$Index),] # sort

### get gene count data only
Gene.abund <- Gene.corr.hier.abund[,(1+ncol(Gene.corr.hier)):ncol(Gene.corr.hier.abund)]
  row.names(Gene.abund)<-Gene.corr.hier.abund$Gene # make genes rownames, before transfo
rm
```

Then normalize relative abundance of genes across sites (z-scores) and gather gene and class data for interpretation

```
# transpose first so normalized by relative abundance in each soil, transpose back
Gene.abundZ<-scale(t(log(Gene.abund)), center=TRUE, scale=TRUE)  # LOG counts before sca
le to norm. distrib.
  Gene.abundZ<-data.matrix(t(Gene.abundZ)) #  transpose again to rotate fig

# Replace gene ID with ID + short names
row.names(Gene.abundZ) <-Gene.corr.hier.abund$Gene_name
# row.names(Gene.abundZ) <-Gene.corr.hier.abund$Gene_fxn  # or full fxn names

# Get colors for labeling soil C, gene class and categories
Gene.class.color<-as.vector(unlist(Gene.corr.hier.abund$Class_color))
Gene.category.color<-as.vector(unlist(Gene.corr.hier.abund$Category_color))
Field.C.color<-as.vector(unlist(soil.chem$Field_C_color))
```

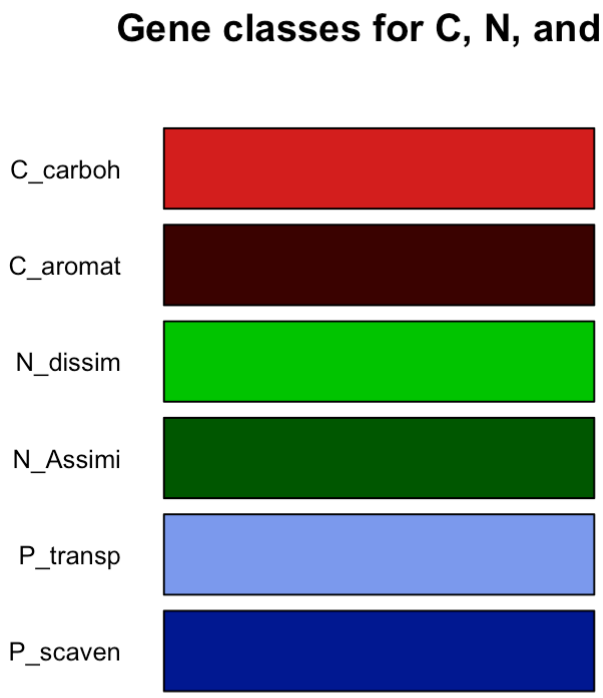
c) Display colors legends for gene categories and soils (heatmap row and columns)

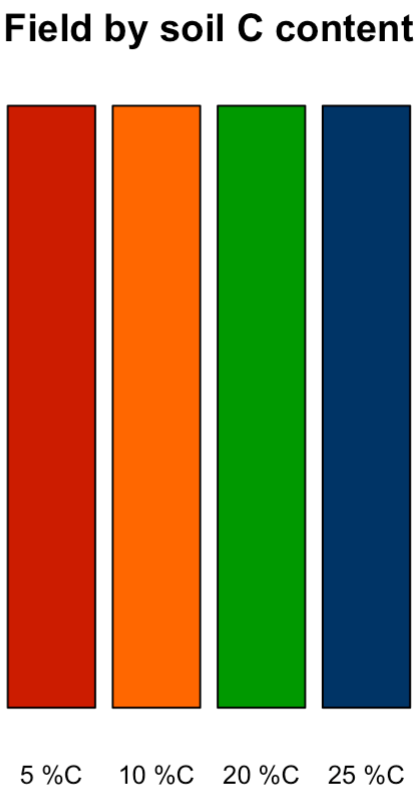
```
par(mfrow=c(1,2))

# Make color legends for functional groups of genes
  Gene.class.levels <- as.character(substr(unique(Gene.corr.hier.abund$Class),1,8))
  Gene.class.counts <- c(rep(1,length(Gene.class.levels)))
  Gene.class.colors <- as.character(unique(Gene.corr.hier.abund$Class_color))
barplot(Gene.class.counts, col=rev(Gene.class.colors), names.arg=rev(Gene.class.levels),

        axes=FALSE, las=2, cex.names=0.8, horiz=TRUE, main="Gene classes for C, N, and
 P")

# Make color legends for soils
  Soil.levels <-as.character(unique(soil.chem$Field_C))
  Soil.counts <- c(rep(1,length(Soil.levels)))
  Soil.colors <- as.character(unique(soil.chem$Field_C_color))
barplot(Soil.counts, col=Soil.colors, names.arg=Soil.levels, axes=FALSE,
        main = "Field by soil C content",cex.names=0.8)
```

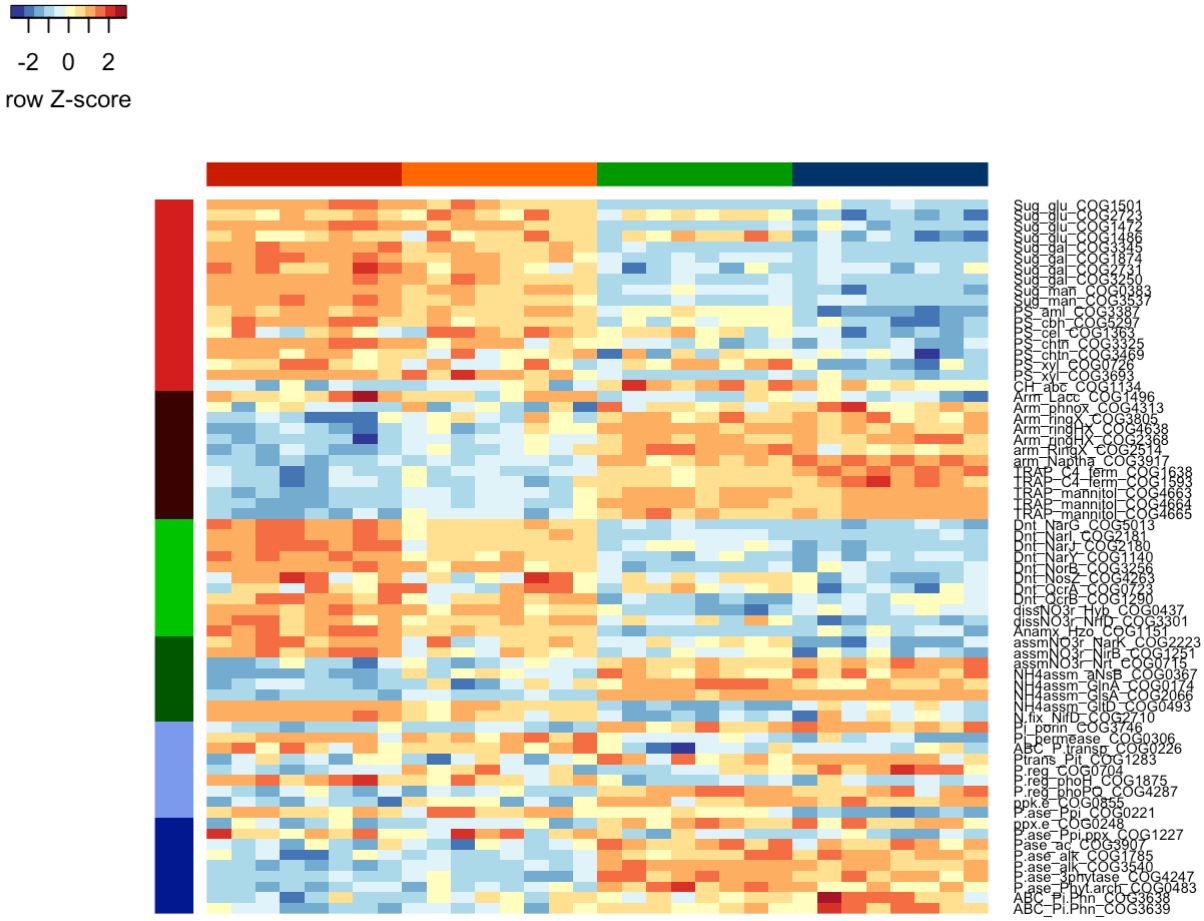





```
par(mfrow=c(1,1)) # return graphics state to original config
```

d) Plot heatmap of correlated genes by function and soils

```
heatmap.2(Gene.abundZ, Rowv=F, Colv=F, RowSideColors=Gene.class.color,
          ColSideColors=Field.C.color, col = rev(brewer.pal(11,"RdYlBu")),
          trace="none", key=TRUE, density.info="none",
          keysize=1, key.title = NA, key.xlab="row Z-score",
          margins = c(1, 10), cexRow=0.7, cexCol=0.5)
```



In order to see how well the data groups itself by gene categories and soil types, I did not using typical hierarchical clustering of genes or sites here. The groupings by category are already pretty good!

e) Abundance of aggregated gene categories for C, N, and P vary with soil N:P ratios

Since these groupings look so nice in the heatmap, why not plot them together as a group?

```
# get functional gene classes and aggregate gene counts
Class<-Gene.corr.hier.abund$Class
Aggr.CNP.fxns <- aggregate(Gene.abund, by=list(Class), FUN=sum) # aggregate
  row.names(Aggr.CNP.fxns) <-Aggr.CNP.fxns[,1] # add row.names before t
Agg.CNP.fxns<-t(Aggr.CNP.fxns[,-1]) # transpose

# combine with soil N:P ratio data
Soil.N_P <-data.frame("N_P"= soil.chem$N_P) # get soil N:P
Agg.CNP.DS <-cbind(Soil.N_P, Agg.CNP.fxns)  # join data
```

Plot genes classes vs. soil N:P, showing Redfield ratio (N:P = 16:1)

```
# Carbon gene classes
Ccp <-ggplot(Agg.CNP.DS, aes(x=N_P, y=C_carbohydrates)) +
  geom_point(color=Gene.class.colors[1]) + scale_x_log10(breaks=c(12,16,20,30,50,70)) +
  geom_smooth(method="auto", se=FALSE, fullrange=FALSE, color=Gene.class.colors[1],linet
ype="solid") +
  geom_vline(xintercept=16, linetype="dashed") +
  labs(title="Carbon", x="Soil N:P", y= "Carbohyd. genes")

Cap <-ggplot(Agg.CNP.DS, aes(x=N_P, y=C_aromatics)) +
  geom_point(color=Gene.class.colors[2]) + scale_x_log10(breaks=c(12,16,20,30,50,70)) +
  geom_smooth(method="auto", se=FALSE, fullrange=FALSE, color=Gene.class.colors[2],linet
ype="solid") +
  geom_vline(xintercept=16, linetype="dashed") +
  labs(x="Soil N:P", y= "Aromatic genes")
```

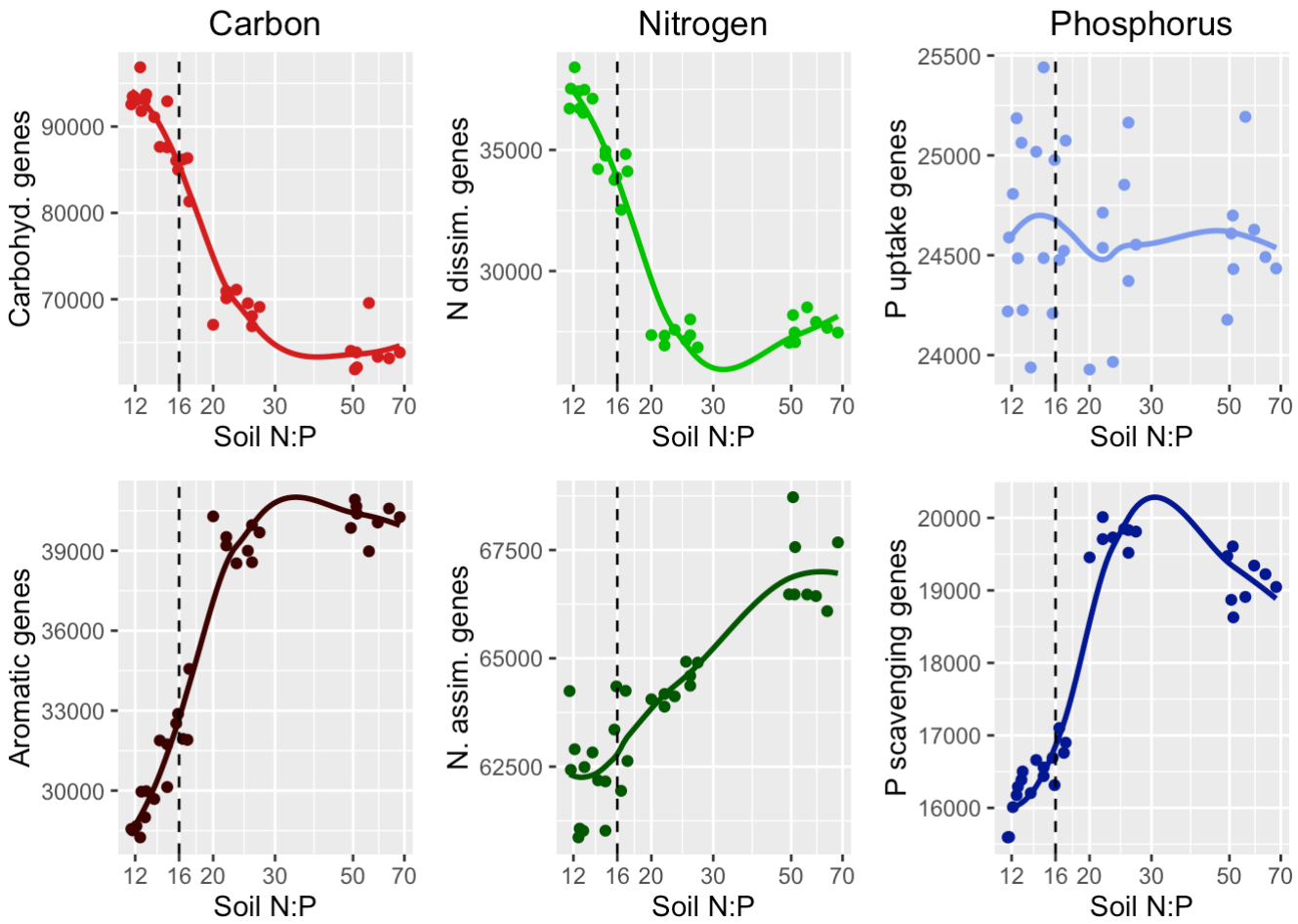
```
# Nitrogen gene classes
Ndp <-ggplot(Agg.CNP.DS, aes(x=N_P, y=N_dissimilation)) +
  geom_point(color=Gene.class.colors[3]) + scale_x_log10(breaks=c(12,16,20,30,50,70)) +
  geom_smooth(method="auto", se=FALSE, fullrange=FALSE, color=Gene.class.colors[3],linet
ype="solid") +
  geom_vline(xintercept=16, linetype="dashed") +
  labs(title="Nitrogen", x="Soil N:P", y= "N dissim. genes")

Nap <-ggplot(Agg.CNP.DS, aes(x=N_P, y=N_Assimilation)) +
  geom_point(color=Gene.class.colors[4]) + scale_x_log10(breaks=c(12,16,20,30,50,70)) +
  geom_smooth(method="auto", se=FALSE, fullrange=FALSE, color=Gene.class.colors[4],linet
ype="solid") +
  geom_vline(xintercept=16, linetype="dashed") +
  labs(x="Soil N:P", y= "N. assim. genes")
```

```
# Phosphorus gene classes
Ptp <-ggplot(Agg.CNP.DS, aes(x=N_P, y=P_transp_intnl)) +
  geom_point(color=Gene.class.colors[5]) + scale_x_log10(breaks=c(12,16,20,30,50,70)) +
  geom_smooth(method="auto", se=FALSE, fullrange=FALSE, color=Gene.class.colors[5],linet
ype="solid") +
  geom_vline(xintercept=16, linetype="dashed") +
  labs(title="Phosphorus", x="Soil N:P", y= "P uptake genes")

Psp <-ggplot(Agg.CNP.DS, aes(x=N_P, y=P_scavenging)) +
  geom_point(color=Gene.class.colors[6]) + scale_x_log10(breaks=c(12,16,20,30,50,70)) +
  geom_smooth(method="auto", se=FALSE, fullrange=FALSE, color=Gene.class.colors[6],linet
ype="solid") +
  geom_vline(xintercept=16, linetype="dashed") +
  labs(x="Soil N:P", y= "P scavenging genes")

grid.arrange(Ccp, Ndp, Ptp, Cap, Nap, Psp,ncol=3, nrow =2)
```



Microbes in lower % soil carbon (C) fields have greater potential carbohydrate utilization, while higher % C fields use more aromatic compounds. Also, lower % C fields have more potential dissimilatory nitrogen N metabolism (e.g denitrification), while higher % C fields have greater P scavenging, consistent with their higher N:P ratios.

n.b. I have used a less refined classification scheme for P uptake genes here, which accounts for differences in P uptake gene patterns compared to the manuscript. Here P uptake genes include many regulators and internal cycling like polyphosphate kinase. Note the noisiness of these genes in the heatmap. Also, here the y values are counts per 10 M, not per 1M as in the paper.

Better regression fitting methods were used in the paper, comparing Box-Cox and polynomial regressions.