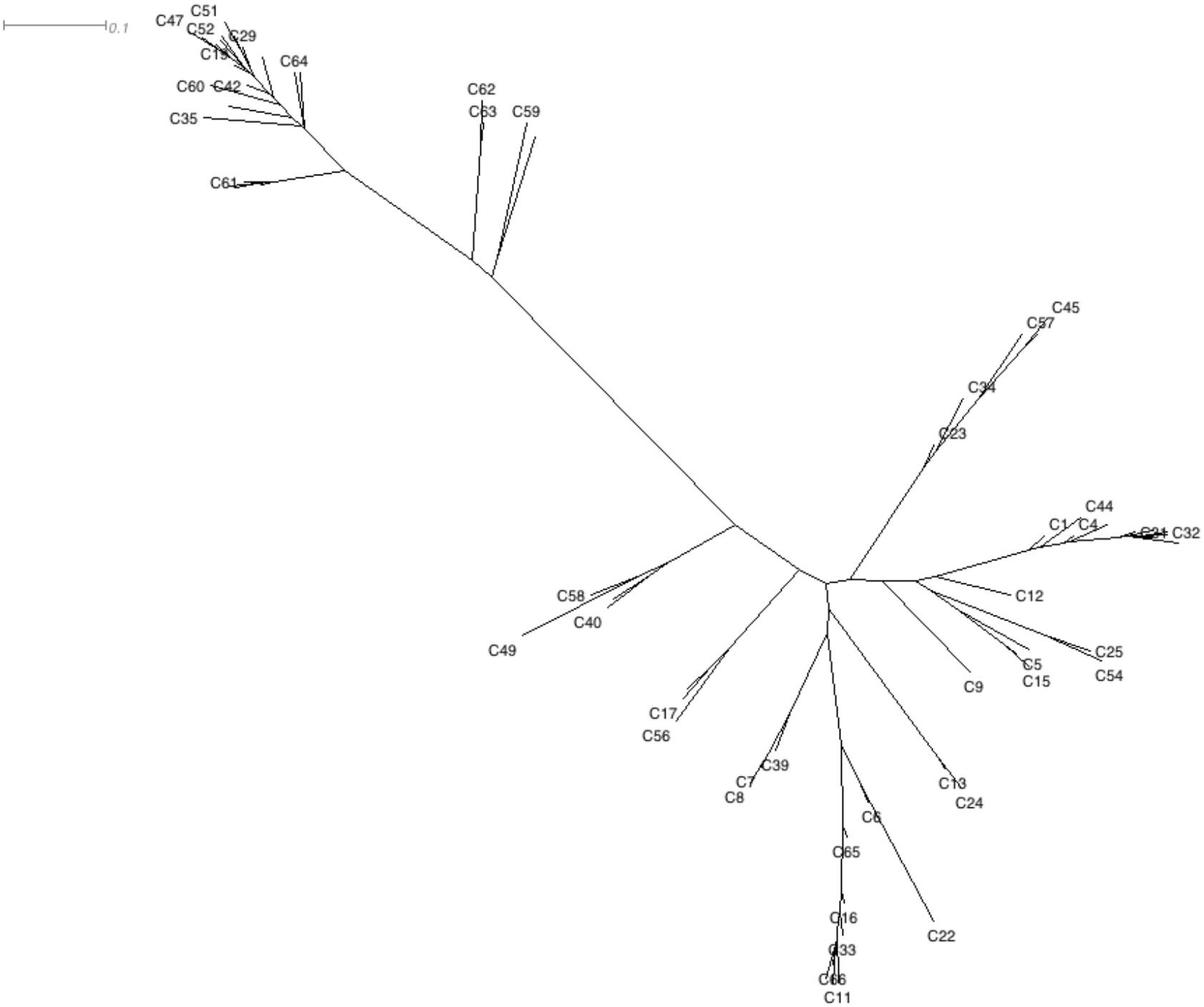
Gradients in Microbial Community Analysis

Christopher Quince

Metapop NERC 2014

Introduction

- Gradients are highly important in structuring microbial communities
- Examine one example data set comprising archaeal amoA gene from 46 soils “Niche specialization of terrestrial archaeal ammonia oxidizers “ (Gubry-Rangin et al. PNAS 2012)
- Protein coding interesting implications for noise removal
- 592 bp amplicons assembled via pairwise comparisons of forward and reverse reads
- 67 5% similarity average linkage OTUs


0.1
C47
C51
C52
C29
C15
C64
C60
C42
C35
C61
C62
C63
C59
C45
C57
C34
C23
C44
C1
C4
C31
C32
C12
C58
C40
C49
C25
C54
C5
C15
C9
C17
C56
C39
C7
C8
C13
C24
C6
C65
C16
C22
C33
C66
C11

Installing R

- R can be downloaded from:

  http://www.r-project.org

  There are pre-compiled binaries available for Windows and Mac

  Answers to frequently asked questions about R are available here:

  http://cran.r-project.org/doc/FAQ/R-FAQ.html

  http://cran.r-project.org/bin/windows/base/rw-FAQ.html (FAQ on R for Windows)

  There is a good introduction to R here:

  http://cran.r-project.org/doc/manuals/R-intro.html
- For this session, you can use R on your amazon cloud EC2 image
  - Red commands to run

Getting started on the EC2

- Logon to amazon cloud and start up a terminal
- Get the tutorial from my Public Dropbox:

  wget https://dl.dropboxusercontent.com/u/7163977/MultivariateStats.tar.gz
- Go into Tutorials, expand directory and move into it:

  tar –xvzf MultivariateStats.tar.gz

  cd MultivariateStats

Importing data and loading libraries

To start R command line on server type R. Type the commands in red at the R command line. Do not include the initial ">". You can redisplay and edit previous commands using the arrow keys

- Import data:

  >AS_C05 <- read.csv("AllSites_C05.csv",header=TRUE,row.names=1)

  >Env <- read.csv("Env.csv",header=TRUE,row.names=1)

  >pH <- Env$pH

- Install libraries not all necessary:

  >install.packages("mgcv")

  >install.packages("picante")

  >install.packages("gplots")

  >install.packages("ggplot2")

  >install.packages("RColorBrewer")

  >install.packages("vegan")

  >install.packages("ape")

  >install.packages("GUniFrac")

- Load libraries:

  >library("mgcv")

  >library("picante")

  >library("gplots")

  >library("ggplot2")

  >library("RColorBrewer")

  >library("vegan")

  >library("ape")

  >library("GUniFrac")

Species Richness

- Sample sizes and species richness:

  >AS <- t(AS_C05)

  >N <- rowSums(AS)

  >S <- specnumber(AS)

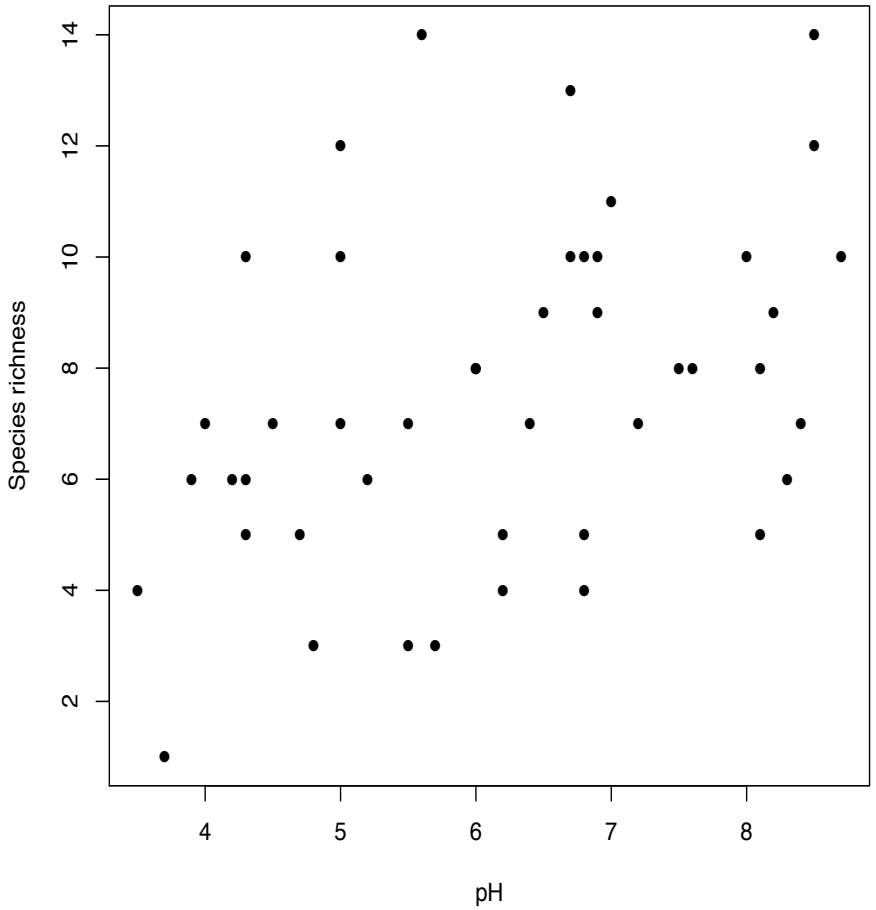
- Is species richness related to pH?

  > qplot(pH,S, geom=c("smooth","poi nt"))+ xlab("pH") + ylab("Species richness”)

- Is it significant?

  > cor.test(pH,S)

- Yes at p = 0.005%



Species Richness (cont.)

- but should rarefy to account for sample size..

  > summary(N)

  >S.rar <- rarefy(AS,482)

  >cor.test(pH,S.rar)
- But now p = 1% …

  >cor.test(pH,N)
- Because N (sample size) and pH are uncorrelated!
- Linear multivariate regression reveals that only pH impacts species richness …

  >S.lm <- lm(S ~ pH + C + N + CN + Moisture + LOI + vegetation, data = Env)

  >summary(S.lm)

Phylogenetic Diversity

- Other diversity measures available e.g. Shannon:

  >Sh <- diversity(AS, index = "shannon", MARGIN = 1, base = exp(1))

- Phylogentic diversity (PD) is a diversity measure that accounts for phylogenetic distance. Normalise frequency matrix and read in tree:

  >ASP <- AS/rowSums(AS)

  >tr <- read.tree("RAxML_bestTree.AllSite.tree")

- Calculate phylogentic diversity, plot, and test for significant relationship with pH (much higher!):

  >pd.result <- pd(ASP, tr, include.root = TRUE)

  >plot(pH,pd.result$PD)

  >cor.test(pH,pd.result$PD)

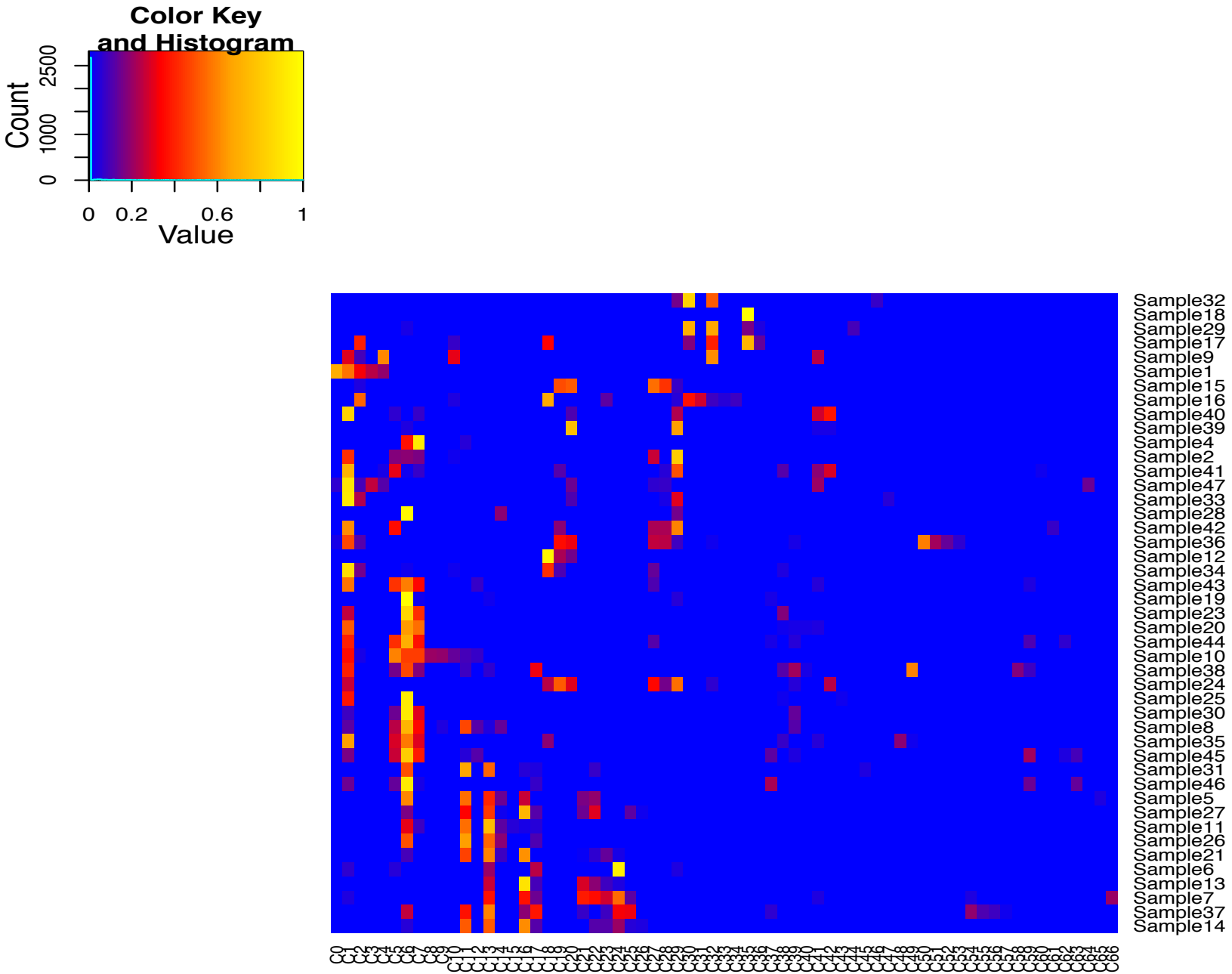
Generating Heat Map

- Make palette and order samples by pH:

```
>crp <-
   colorRampPalette(c("blue","red","orange","ye
   llow"))(100)
>ASPPH <- data.frame(ASP,pH)
>ASPPH.order <- as.matrix(ASPPH[order(pH),])
>ASPO <- ASPPH.order[,1:67]
```

- Plot heat map without reordering

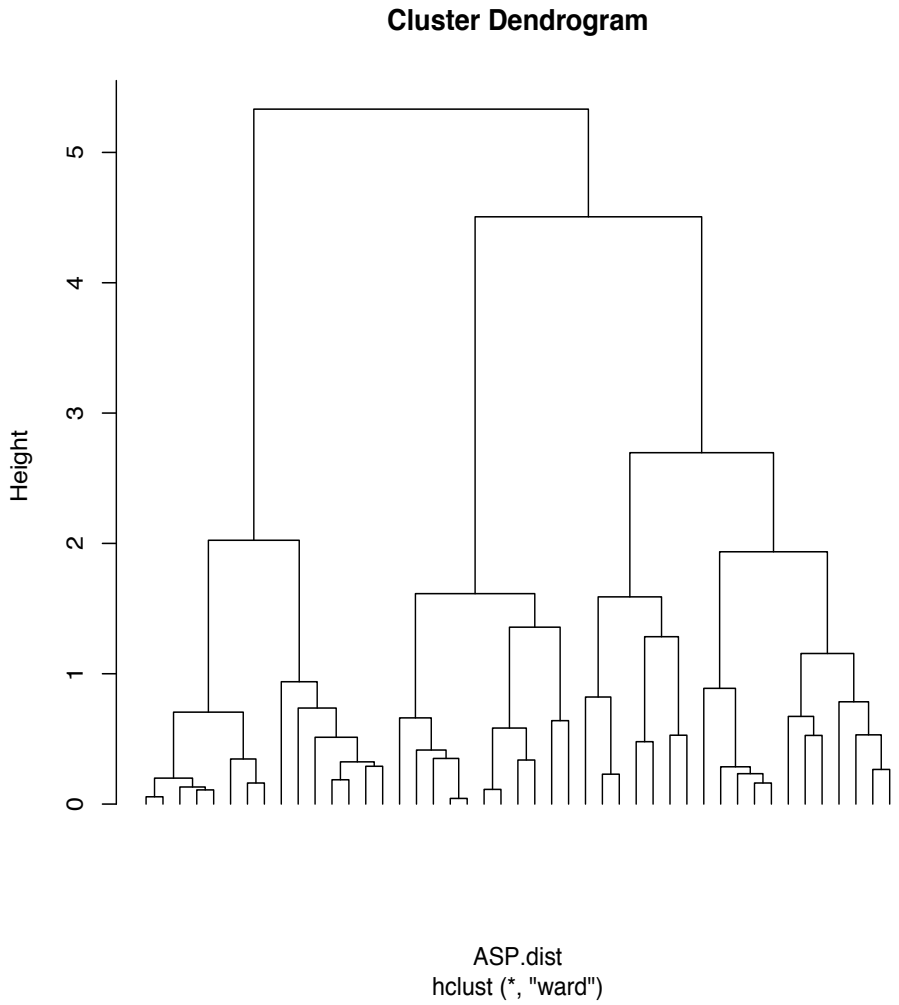
```
>heatmap.2
   (sqrt(ASPO),col=crp,trace="none",Rowv=FALS
   E,Colv=FALSE)
```


Color Key
and Histogram
Count
0
1000
2500
0
0.2
0.6
1
Value
Sample32
Sample18
Sample29
Sample17
Sample9
Sample1
Sample15
Sample16
Sample40
Sample39
Sample4
Sample2
Sample41
Sample47
Sample33
Sample28
Sample42
Sample36
Sample12
Sample34
Sample43
Sample19
Sample23
Sample20
Sample44
Sample10
Sample38
Sample24
Sample25
Sample30
Sample8
Sample35
Sample45
Sample31
Sample46
Sample5
Sample27
Sample11
Sample26
Sample21
Sample6
Sample13
Sample7
Sample37
Sample14

Hierarchical Clustering

- Generate sample distance matrix from relative frequencies:

```
> ASP.dist <-
    vegdist(ASP,dist="bray”
    )
>ASP.hclust <-
    hclust(ASP.dist, method
    = "ward")
>plot(ASP.hclust)
```



MDS using Unifrac

- Calculate Unifrac distances:

```
>ASP.gunifrac <- GUniFrac(ASP, tr, alpha=c(0, 0.5, 1))$unifracs
```

- Extract weighted Unifrac distances:

```
>ASP.uf <- ASP.gunifrac[,,"d_1”]
```

- Perform principle coordinates analysis:

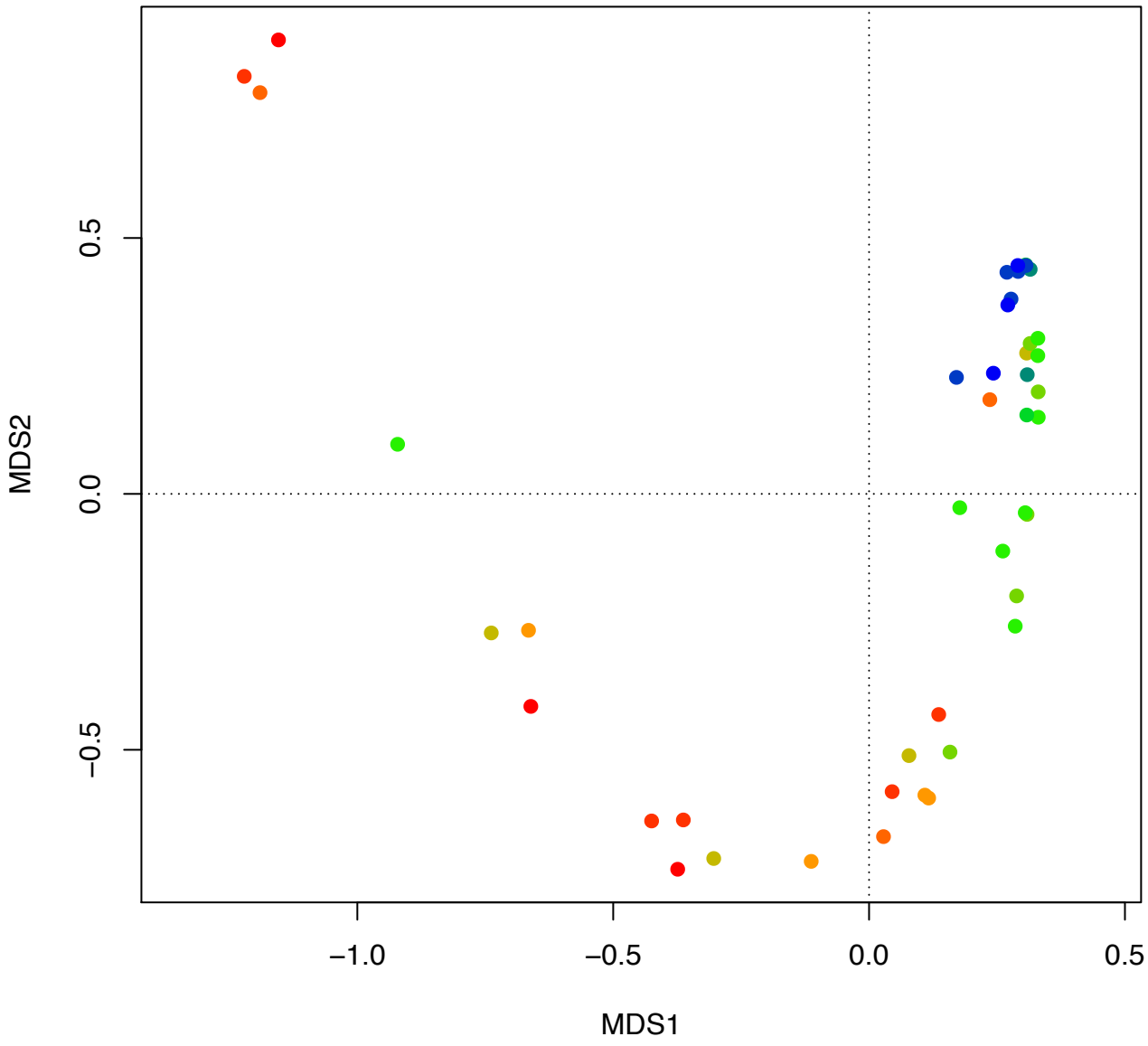
```
>ASP.uf.cap <- capscale(ASP.uf ~ 1)
```

- Rescale pH to integers and make and bind pH like color palette:

```
>IPH <- floor((pH - 3.5)*2) + 1
>crp2 <- colorRampPalette(c("red","orange","green","blue","darkblue"))(14)
>palette(crp2)
```

- Plot:

```
>ordiplot (ASP.uf.cap, display = 'si', type = 'n')
> for (i in seq (1, 14)) points (ASP.uf.cap, select = (IPH == i), col = i, pch = 19)
```


MDS2
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
MDS1

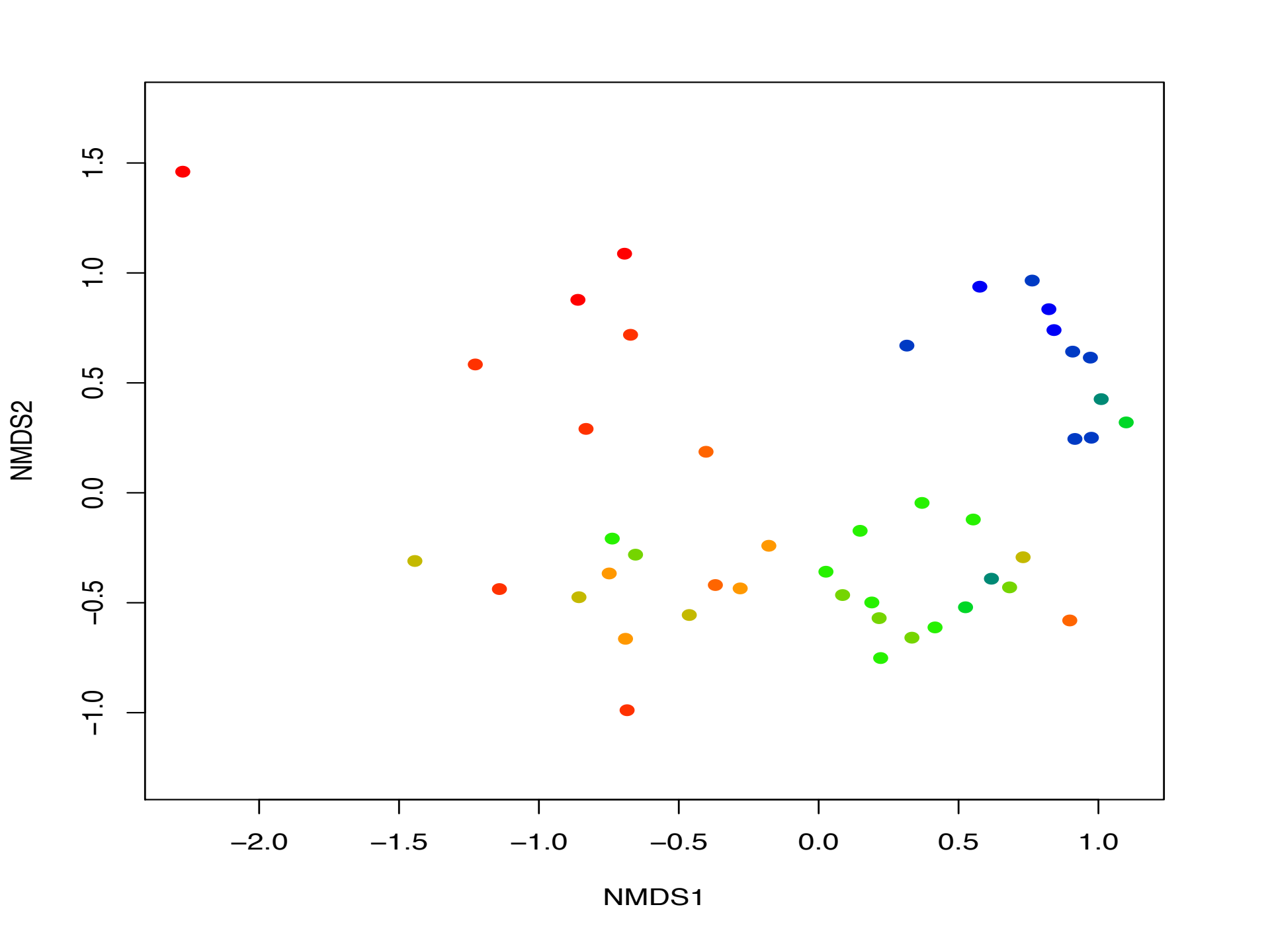
Non-metric Multidimensional Scaling

- Perform NMDS using vegan metaMDS:

```
>ASP.nmds <- metaMDS(ASP)
```

- Plot NMDS empty and add in sites coloured by pH:

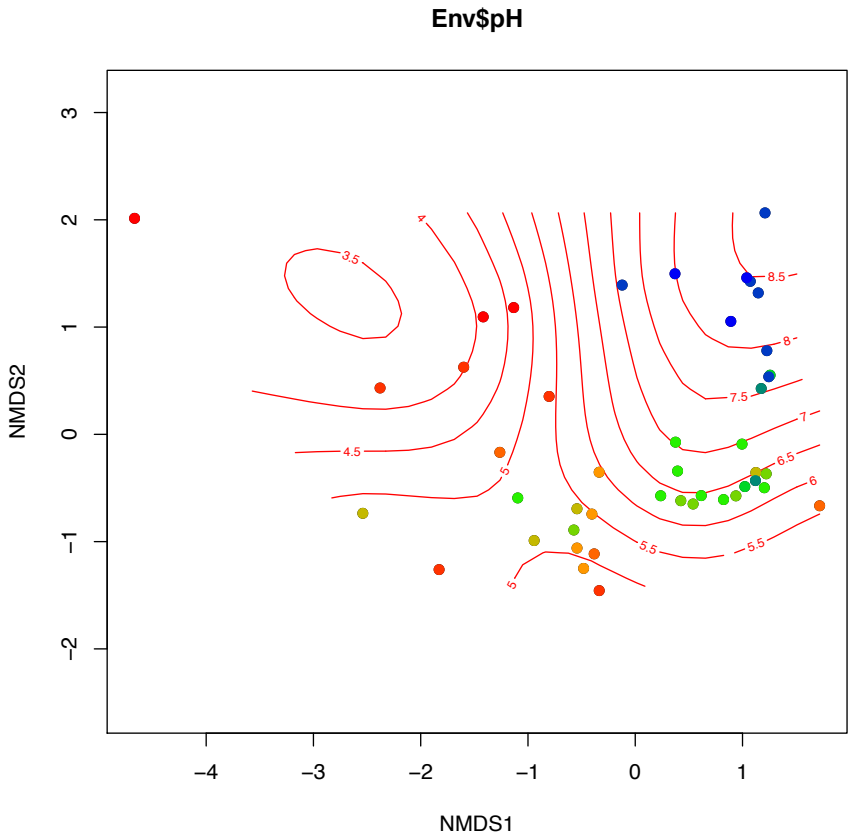
```
> ordiplot (ASP.nmds, display = 'si', type = 'n')
> for (i in seq (1, 14)) points (ASP.nmds, select =
  (IPH == i), col = i, pch = 19)
```


NMDS2
1.5
1.0
0.5
0.0
−0.5
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
NMDS1

Adding pH gradient…

- Very easy to do:

```
>ordisurf(ASP.nmds, Env
    $pH)
>for (i in seq (1, 14))
    points (ASP.nmds,
    select = (IPH == i), col =
    i, pch = 19)
```



NMDS Using Phylogentic Distance Metric (MPD)

- First need to generate cophenetic distance matrix from tree:

```
>tr.phydist <- cophenetic(tr)
```

- Use this to calculate mean pairwise distance between all communities:

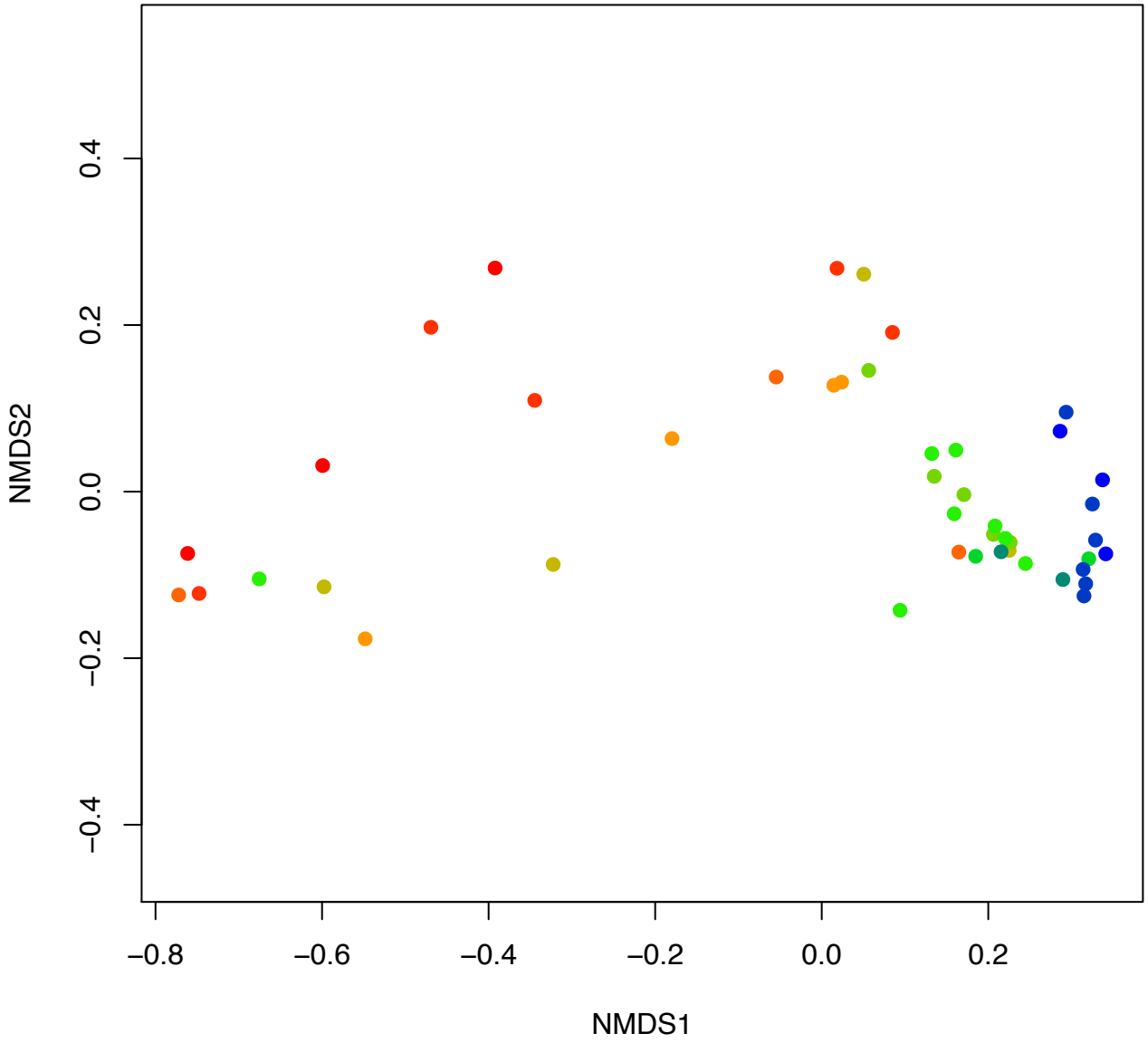
```
>ASP.comdist <- comdist(ASP, tr.phydist,abundance.weighted=TRUE)
```

- Perform NMDS using vegan metaMDS on those distances:

```
>ASP.comdist.nmds <- metaMDS(ASP.comdist)
```

- Plot NMDS empty and add in sites coloured by pH:

```
> ordiplot (ASP.comdist.nmds, display = 'si', type = 'n')
> for (i in seq (1, 14)) points (ASP.comdist.nmds, select = (IPH == i), col
    = i, pch = 19)
```


NMDS2
0.4
0.2
0.0
−0.2
−0.4
−0.8
−0.6
−0.4
−0.2
0.0
0.2
NMDS1

Correspondence Analysis

- Long gradient suggests correspondence rather than redundancy analysis:

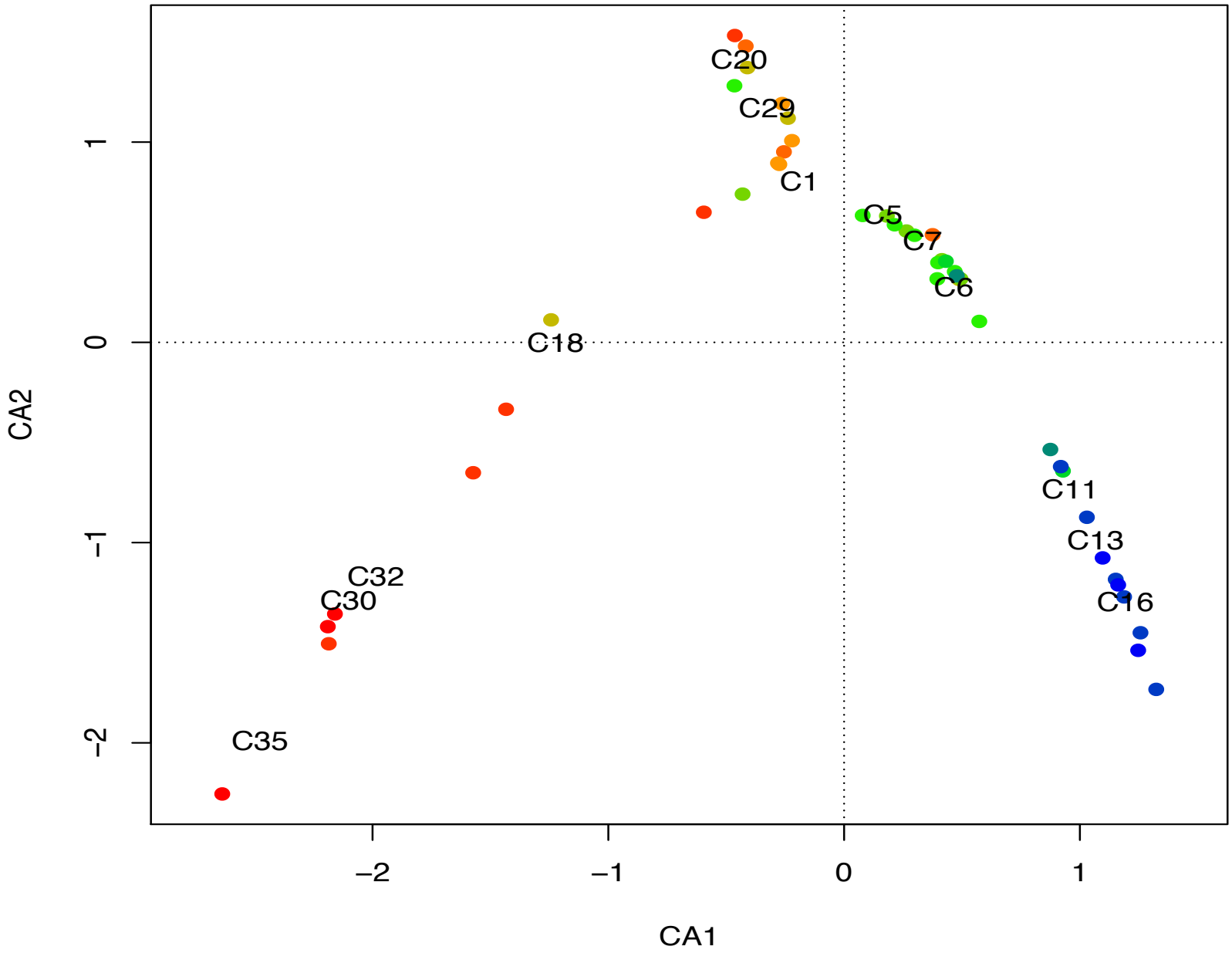
```
>ASP.ca <- cca(ASP)
```

- Select species with over 3,000 reads:

```
>selSp <- colSums(AS)>3000
```

- Generate plot:

```
> ordiplot (ASP.ca, display = 'si', type = 'n')
> for (i in seq (1, 14)) points (ASP.ca, select = (IPH == i), col = i, pch = 19)
> text(ASP.ca,display='sp',select=selSp)
```


C20
C29
C1
C5
C7
C6
C18
C11
C13
C16
C32
C30
C35
CA1
CA2
−2
−1
0
1
−2
−1
0
1

Canonical Correspondence Analysis

- Use same cca function but include regression formula:

  >ASP.cca <- cca(ASP ~ pH + CN + LOI + Moisture+ vegetation,data=Env)
- What about significance – use random permutations of columns (OTUs) of community matrix?

  >anova(ASP.cca)

  >anova(ASP.cca, by="terms”)

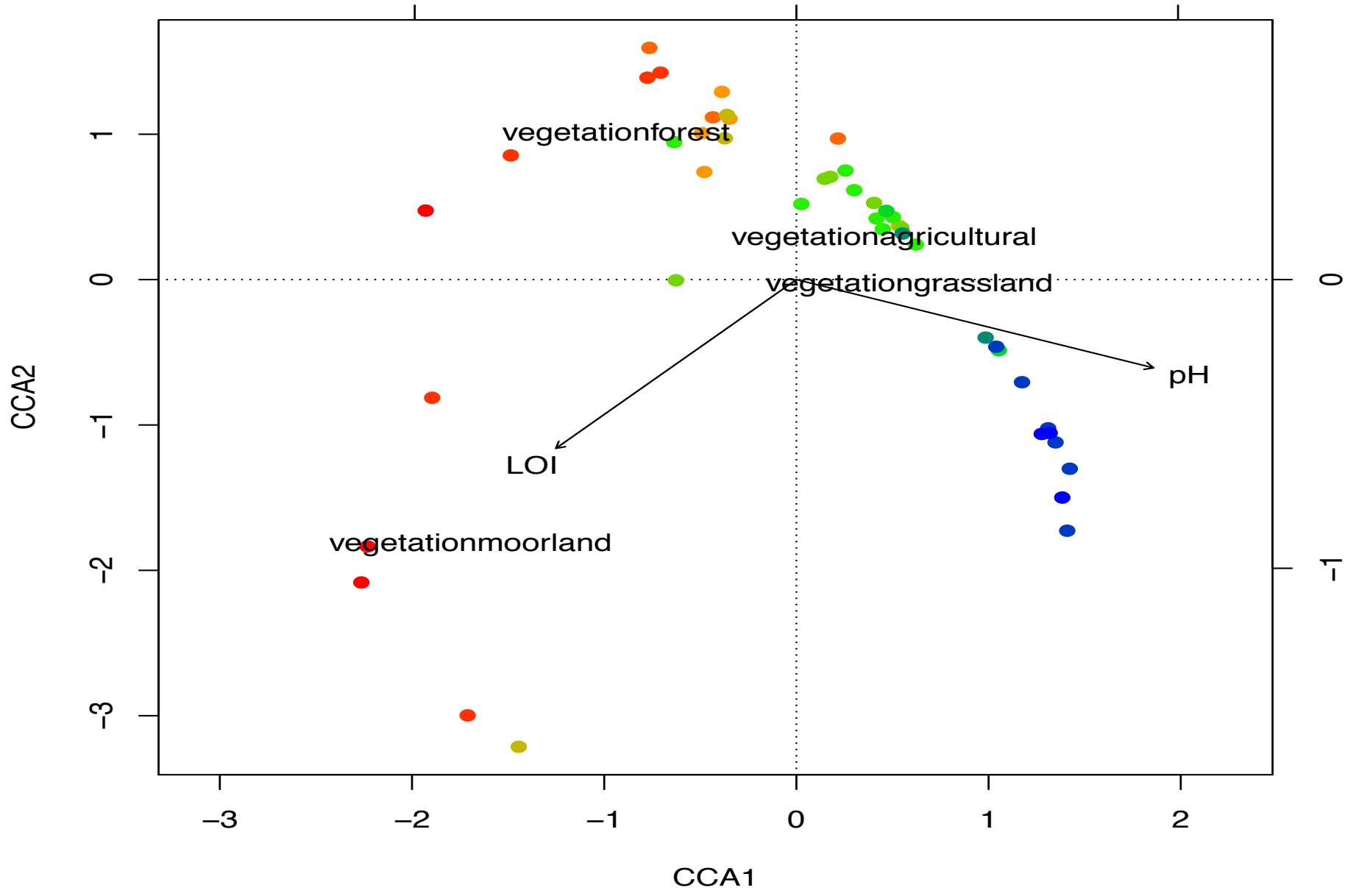
  >ASP.cca <- cca(ASP ~ pH + CN + LOI + Moisture+ vegetation,data=Env)
- In original, reference cluster study, only pH significant, now find pH**, LOI** and vegetation*. Redo CCA with these and generate plot:

  >ASPR.cca <- cca(ASP ~ pH + LOI + vegetation,data=Env)

  > ordiplot(ASPR.cca, display = 'si', type = 'n')

  > for (i in seq (1, 14)) points (ASPR.cca, select = (IPH == i), col = i, pch = 19)

  > text(ASPR.cca,"cn")


vegetationforest
vegetationagricultural
vegetationgrassland
pH
LOI
vegetationmoorland
CCA1
CCA2
−3
−2
−1
0
1
2
−3
−2
−1
0
1
0
−1

Principal coordinates

- Use same cca function but include regression formula try with Bray-Curtis, MPD and Unifrac:

  >ASP.cap <- capscale(ASP ~ .,data=Env)

  >ASP.comdist.cap <- capscale(ASP.comdist ~ .,data=Env)

  >ASP.uf.cap <- capscale(ASP.uf ~ .,data=Env)
- What about significance – use random permutations of columns (OTUs) of community matrix?

  >anova(ASP.comdist.cap)

  >anova(ASP.comdist.cap, perm.max=2000,perm=2000,by="terms")

  >anova(ASP.uf.cap, perm.max=2000,perm=2000,by="terms")

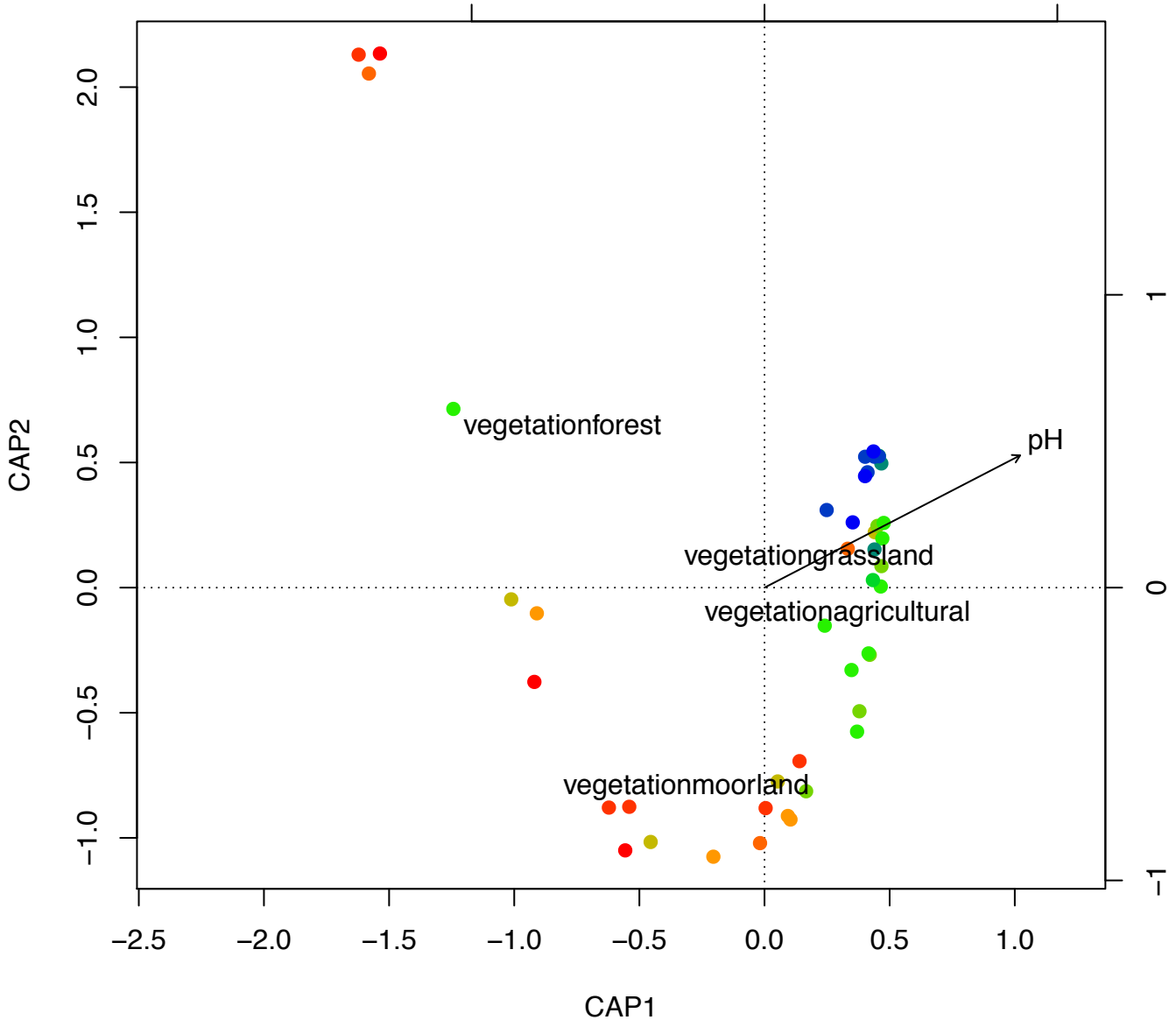
  >anova(ASP.cap, perm.max=2000,perm=2000,by="terms")
- For Unifrac pH and vegation lets plot ordination with these:

  >ASPR.uf.cap <- capscale(ASP.uf ~ pH + vegetation,data=Env)

  >ordiplot(ASPR.uf.cap, display = 'si', type = 'n')

  >for (i in seq (1, 14)) points (ASPR.uf.cap, select = (IPH == i), col = i, pch = 19)

  >text(ASPR.uf.cap,"cn")


vegetationforest
pH
vegetationgrassland
vegetationagricultural
vegetationmoorland
CAP1
CAP2
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
2.0
1.5
1
0
−1

Hypothesis testing without ordination

- Permutational Multivariate Analysis apply to any model e.g. bray-curtis:

  >ASP.ado <- adonis(ASP ~ ., data=Env)

  >ASP.ado
- Compare to phylogenetically aware metric:

  >ASP.comdist.ado <- adonis(ASP.comdist ~ ., data=Env)

  >ASP.comdist.ado
- And Unifrac:

  >ASP.uf.ado <- adonis(ASP.uf ~ ., data=Env)

  >ASP.uf.ado

We can also do Mantel tests

- Can only account dissimilarity matrix for continuous environmental variables:

```
>EnvN <- Env[,1:6]
>EnvN.dist <- vegdist(scale(EnvN), "euclid")
>mantel(ASP.dist,EnvN.dist)
>mantel(ASP.uf,EnvN.dist)
```

Relationship of the most abundant groups to pH

- Sort OTU total frequencies:

```
>sort(colSums(AS))
```

- Log-transform normalised OUT frequencies with pseudo-count:

```
>logASP <- log((AS + 1)/rowSums(AS + 1))
```

- Pull out relative frequencies of three most abundant + C30:

```
>logC6 <- logASP[,"C6"]
>logC1 <- logASP[,"C1"]
>logC13 <- logASP[,"C13"]
>logC30 <- logASP[,"C30"]
```

- Use penalized generalized additive model to fit to relative frequencies:

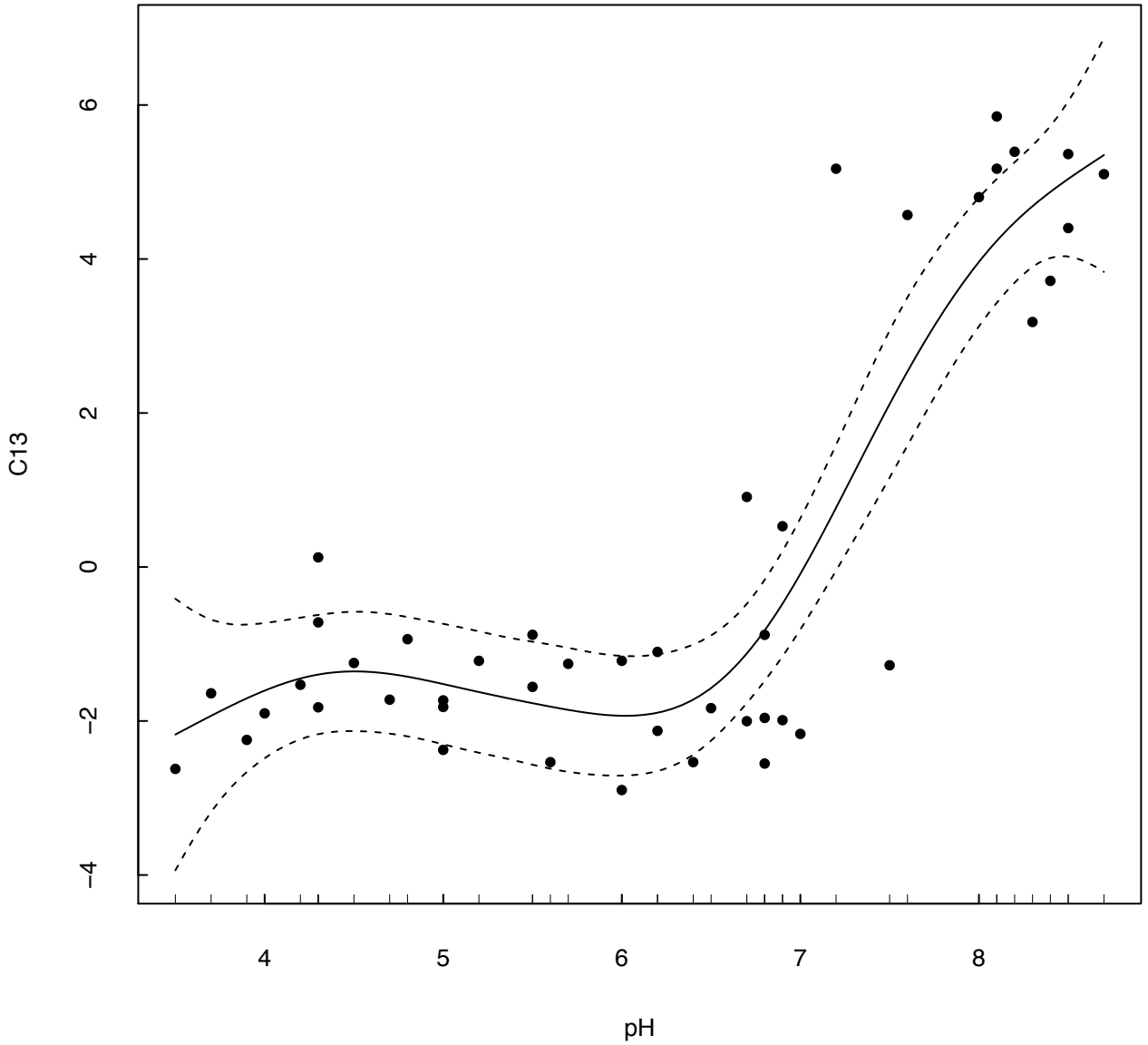
```
>logC6.gam<-gam(logC6~s(pH))
>summary(logC6.gam)
```

- Highly significant and explain large percentage of variance, plot three fits:

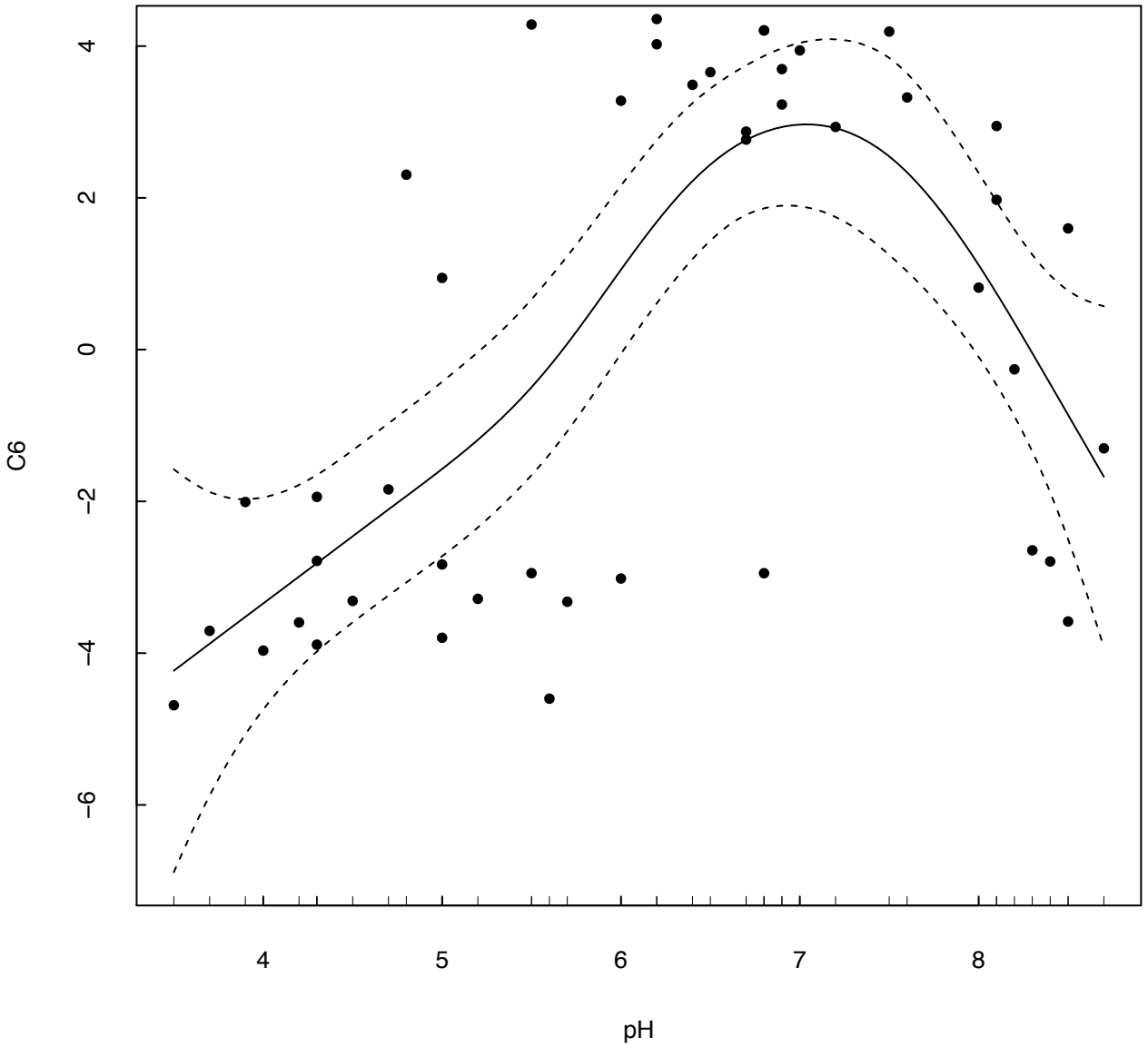
```
>plot(logC6.gam, xlab = "pH", ylab = "C6", las=0, pch=20, cex.axis=0.8,
    tck=0.01, cex.lab=0.85)
>points(pH,logC6 – mean(logC6),pch=20)
```

- Repeat for C1, C13 and C30 if you want

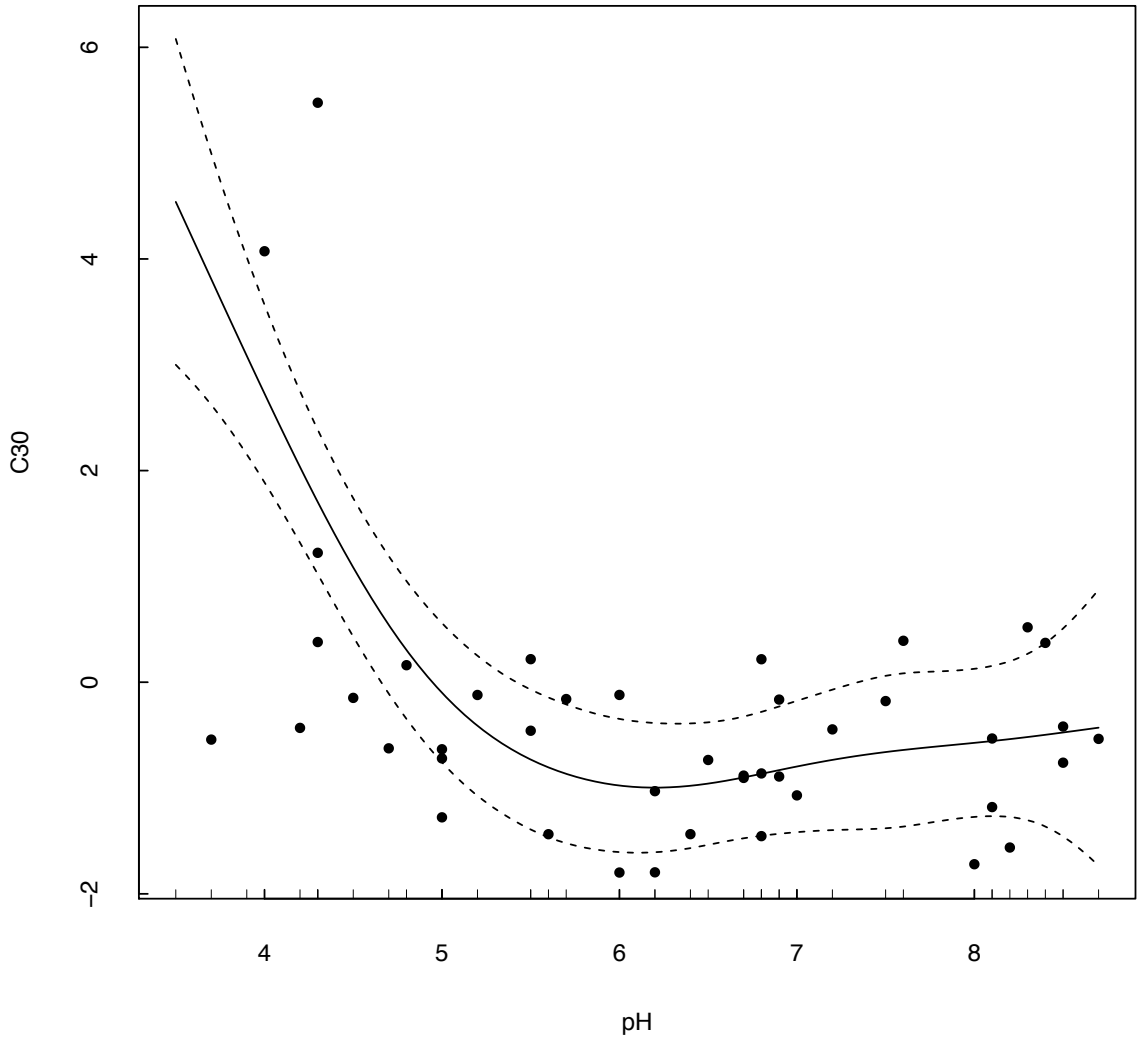
C13 - alkalinophile



C6 – neutralophile



C30 – extreme acidophile



Bonferroni-Hochberg Correction

- To correct for multiple comparisons:

```
nT <- ncol(logASP)
p <- rep(0,nT)
for(i in 1:nT){
    temp <-gam(logASP[,i]~s(pH))
     stemp <- summary(temp)
     p[i] <- stemp$p.table[[4]]
}
pa <- p.adjust(p, method = "BH")
hcp.df <- data.frame(colnames(logASP))
hcp.df <- cbind(hcp.df,p,pa)
head(hcp.df[order(hcp.df$p),],10)
```

Conclusion

- Archael ammonia oxidiser community strongly structured by pH with different OTUs having clear pH range
- Community composition is further differentiated between moorland and forest, grassland and agricultural continuum at 5%