Dormancy and dispersal structure bacterial communities across ecosystem boundaries

Nathan I. Wisnoski, Mario E. Muscarella, Megan L. Larsen, and Jay T. Lennon

24 January, 2019

Initial Setup

First, we'll load the packages we'll need for the analysis, as well as some other functions.

```
# Import Required Packages
library("png")
library("grid")
library("tidyverse")
library("vegan")
library("xtable")
library("viridis")
library("cowplot")
library("adespatial")
library("ggrepel")
library("gganimate")
library("maps")
library("rgdal")
library("iNEXT")
library("officer")
library("flextable") #must have gdtools installed also
library("broom")
library("ggpmisc")
library("pander")

source("bin/mothur_tools.R")
se <- function(x, ...){sd(x, na.rm = TRUE)/sqrt(length(na.omit(x)))}
```

Next, we'll set the aesthetics of the figures we will produce.

```
my.cols <- RColorBrewer::brewer.pal(n = 4, name = "Greys")[3:4]

# Set theme for figures in the paper
theme_set(theme_classic() +
  theme(axis.title = element_text(size = 20),
        axis.title.x = element_text(margin = margin(t = 15, b = 15)),
        axis.title.y = element_text(margin = margin(l = 15, r = 15)),
        axis.text = element_text(size = 14),
        axis.text.x = element_text(margin = margin(t = 5)),
        axis.text.y = element_text(margin = margin(r = 5)),
        #axis.line.x = element_line(size = 1),
        #axis.line.y = element_line(size = 1),
        axis.line.x = element_blank(),
        axis.line.y = element_blank(),
        axis.ticks.x = element_line(size = 1),
        axis.ticks.y = element_line(size = 1),
```

```
        axis.ticks.length = unit(.1, "in"),
        panel.border = element_rect(color = "black", fill = NA, size = 1.5),
        legend.title = element_blank(),
        legend.text = element_text(size = 16),
        strip.text = element_text(size = 14),
        strip.background = element_blank()
        ))
```

Import Data

Here, we read in the processed sequence files from mothur (shared and taxonomy) and a design of the sampling. We also load in the environmental data. We then remove the mock community from the dataset and ensure the the design and OTU table are aligned by row.

```
# Define Inputs
# Design = general design file for experiment
# shared = OTU table from mothur with sequence similarity clustering
# Taxonomy = Taxonomic information for each OTU
design <- "data/UL.design.txt"
shared <- "data/ul_resgrad.trim.contigs.good.unique.good.filter.unique.precluster.pick.pick.pick.opti_m
taxon  <- "data/ul_resgrad.trim.contigs.good.unique.good.filter.unique.precluster.pick.pick.pick.opti_m

# Import Design
design <- read.delim(design, header=T, row.names=1)

# Import Shared Files
OTUs <- read.otu(shared = shared, cutoff = "0.03")    # 97% Similarity

# Import Taxonomy
OTU.tax <- read.tax(taxonomy = taxon, format = "rdp")

# Load environmental data
env.dat <- read.csv("data/ResGrad_EnvDat.csv", header = TRUE)
env.dat <- env.dat[-16,]

# Subset to just the reservoir gradient sites
OTUs <- OTUs[str_which(rownames(OTUs), "RG"),]
OTUs <- OTUs[-which(rownames(OTUs) == "RGMockComm"),]

# make sure OTU table matches up with design order
OTUs <- OTUs[match(rownames(design), rownames(OTUs)),]
```

Clean and transform OTU table

Here, we remove OTUs with low incidence across sites, we remove any samples with low coverage, and we standardize the OTU table by log-transforming the abundances and relativizing by site.

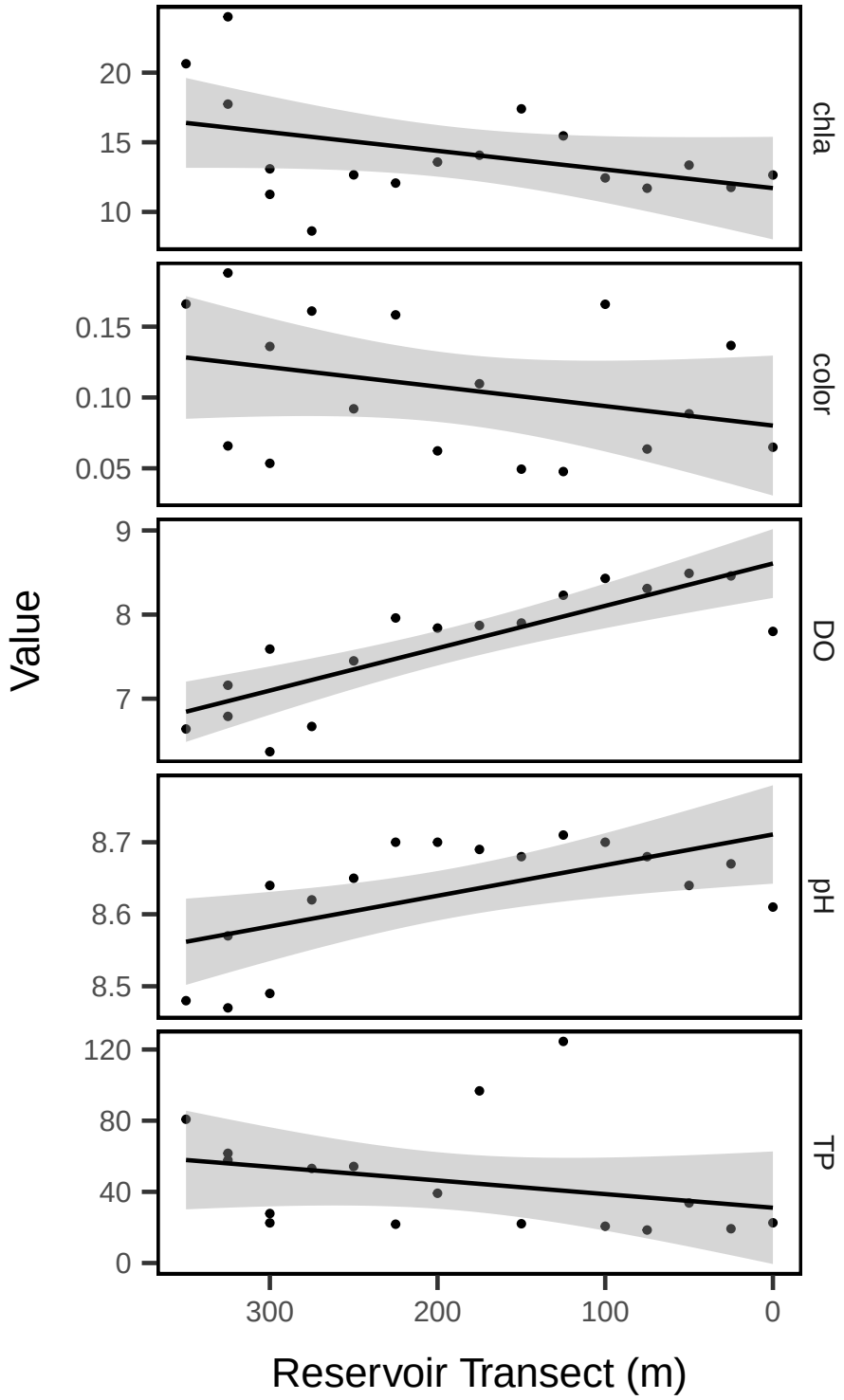
```
# Remove OTUs with less than two occurences across all sites
OTUs <- OTUs[, which(colSums(OTUs) >= 2)]

# Sequencing Coverage
coverage <- rowSums(OTUs)
```

```
# Remove Low Coverage Samples (This code removes two sites: Site 5DNA, Site 6cDNA)
lows <- which(coverage < 10000)
OTUs <- OTUs[-which(coverage < 10000), ]
design <- design[-which(coverage < 10000), ]
# Remove OTUs with less than two occurences across all sites
OTUs <- OTUs[, which(colSums(OTUs) >= 2)]
# OTUs <- rrarefy(OTUs, min(coverage))

# Make Relative Abundance Matrices
OTUsREL <- decostand(OTUs, method = "total")

# Log Transform Relative Abundances
OTUsREL.log <- decostand(OTUs, method = "log")
```

Reservoir environmental gradients

Just to see if there are any strong underlying resource or nutrient gradients in the reservoir, we'll plot them along the distance of the reservoir.

```
env.dat %>% select(dist.dam, DO, pH, TP, color, chla) %>%
  gather(variable, value, -dist.dam) %>%
  ggplot(aes(x = dist.dam, y = value)) +
  geom_point() +
  geom_smooth(method = "lm", color = "black") +
  facet_grid(variable ~., scales = "free") +
  theme(strip.background = element_blank(), strip.text = element_text(size = 14)) +
  labs(x = "Reservoir Transect (m)",
       y = "Value") +
  scale_x_reverse()
```



So, there are some weak gradients, but nothing too prevailing.

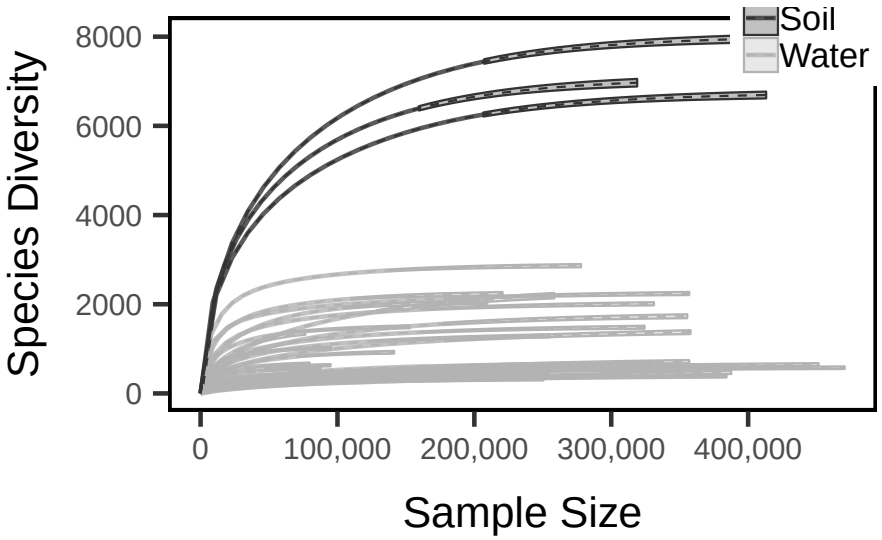
Analyze Diversity

Now, we will analyze the bacterial diversity in the reservoir and nearby soils to figure out how well they support different mechanisms of community assembly.

How does α-diversity vary along the reservoir?

First, we use the method of rarefaction and extrapolation developed by Chao et al. in the iNEXT package.

```
# Observed Richness
S.obs <- rowSums((OTUs > 0) * 1)

# Simpson's Evenness
SimpE <- function(x = ""){
  x <- as.data.frame(x)
  D <- diversity(x, "inv")
  S <- sum((x > 0) * 1)
  E <- (D)/S
  return(E)
}
simpsE <- round(apply(OTUs, 1, SimpE), 3)
shan <- diversity(OTUs, index = "shannon")
exp.shan <- exp(shan)
alpha.div <- cbind(design, S.obs, simpsE, shan, exp.shan)


# # estimate asymptotic richness
# divestim <- iNEXT(t(OTUs), datatype = "abundance", nboot = 999)
# saveRDS(divestim, file = "intermediate-data/inext-output-999boots.rda")
divestim <- read_rds("intermediate-data/inext-output-999boots.rda")
divestim.df <- fortify(divestim) %>%
  mutate(habitat = str_to_title(design[as.character(site),"type"]))

divestim.df %>%
  ggplot(aes(x = x, y = y,
             ymin = y.lwr, ymax = y.upr,
             color = habitat, fill = habitat, group = site)) +
  geom_ribbon(data=subset(divestim.df, method == "extrapolated"), alpha = 0.3) +
  geom_line(data=subset(divestim.df, method == "interpolated"), size = 1, alpha = .8) +
  geom_line(alpha = 1, linetype = "dashed") +
  scale_x_continuous(labels = scales::comma) +
  labs(x = "Sample Size", y = "Species Diversity") +
  theme(legend.position =  c(.9,.95)) +
  scale_color_grey(end = .7) +
  scale_fill_grey(end = .7)
```



Next, we'll extract the estimates for the Hill numbers at different levels of q, which differentially weight common versus rare species.

```
hill.estim <- divestim$AsyEst %>% filter(Diversity == "Species richness") %>%
  left_join(rownames_to_column(alpha.div), by = c("Observed" = "S.obs")) %>%
  select(Site, rowname, station, molecule, type, distance) %>%
  left_join(divestim$AsyEst, by = "Site")

hill.water <- as_tibble(hill.estim) %>% filter(type == "water")
hill.water.rich <- subset(hill.water, Diversity == "Species richness")
hill.water.shan <- subset(hill.water, Diversity == "Shannon diversity")
hill.water.simp <- subset(hill.water, Diversity == "Simpson diversity")

hill.water.mod.rich <- lm(Estimator ~ distance * molecule, data = hill.water.rich)
hill.water.mod.shan <- lm(Estimator ~ distance * molecule, data = hill.water.shan)
hill.water.mod.simp <- lm(Estimator ~ distance * molecule, data = hill.water.simp)

summary(hill.water.mod.rich)
```

```
##
## Call:
## lm(formula = Estimator ~ distance * molecule, data = hill.water.rich)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -518.38 -137.60    0.71   98.61  718.25
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          468.6274   128.2862   3.653 0.000918 ***
## distance               5.2143     0.5397   9.662 5.22e-11 ***
## moleculeRNA          104.9450   166.1694   0.632 0.532164
## distance:moleculeRNA  -5.1222     0.7163  -7.151 4.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 258.7 on 32 degrees of freedom
```

```
## Multiple R-squared:  0.8714, Adjusted R-squared:  0.8593
## F-statistic: 72.26 on 3 and 32 DF,  p-value: 2.431e-14
```

```
summary(hill.water.mod.shan)
```

```
##
## Call:
## lm(formula = Estimator ~ distance * molecule, data = hill.water.shan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.915  -14.841   -3.500    6.902  157.964
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            55.5031    24.9703   2.223  0.03342 *
## distance                0.2892     0.1050   2.753  0.00965 **
## moleculeRNA           -35.5144    32.3441  -1.098  0.28039
## distance:moleculeRNA   -0.2905     0.1394  -2.084  0.04525 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.35 on 32 degrees of freedom
## Multiple R-squared:  0.5556, Adjusted R-squared:  0.514
## F-statistic: 13.34 on 3 and 32 DF,  p-value: 8.138e-06
```

```
summary(hill.water.mod.simp)
```

```
##
## Call:
## lm(formula = Estimator ~ distance * molecule, data = hill.water.simp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.589  -7.222  -0.977   6.321  39.440
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           31.61377    7.45207   4.242 0.000176 ***
## distance               0.05218    0.03135   1.664 0.105804
## moleculeRNA          -22.18812    9.65268  -2.299 0.028205 *
## distance:moleculeRNA  -0.04408    0.04161  -1.059 0.297369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.02 on 32 degrees of freedom
## Multiple R-squared:  0.5693, Adjusted R-squared:  0.529
## F-statistic:  14.1 on 3 and 32 DF,  p-value: 4.985e-06
```

```
hill.water.mods <- as_tibble(rbind.data.frame(
  tidy(hill.water.mod.rich) %>% add_column(Diversity = "Species richness"),
  tidy(hill.water.mod.shan) %>% add_column(Diversity = "Shannon diversity"),
  tidy(hill.water.mod.simp) %>% add_column(Diversity = "Simpson diversity")
))

# Summary table of the model results.
```

```
hill.water.mods %>%
  group_by(Diversity) %>%
  rename("Term" = term,
         "Estimate" = estimate,
         "Std. Error" = std.error,
         "Statistic" = statistic,
         "p-value" = p.value) %>%
  filter(Term != "(Intercept)") %>%
  select(Diversity, everything()) %>%
  mutate_if(is.double, signif, digits = 3) %>%
  pander()
```

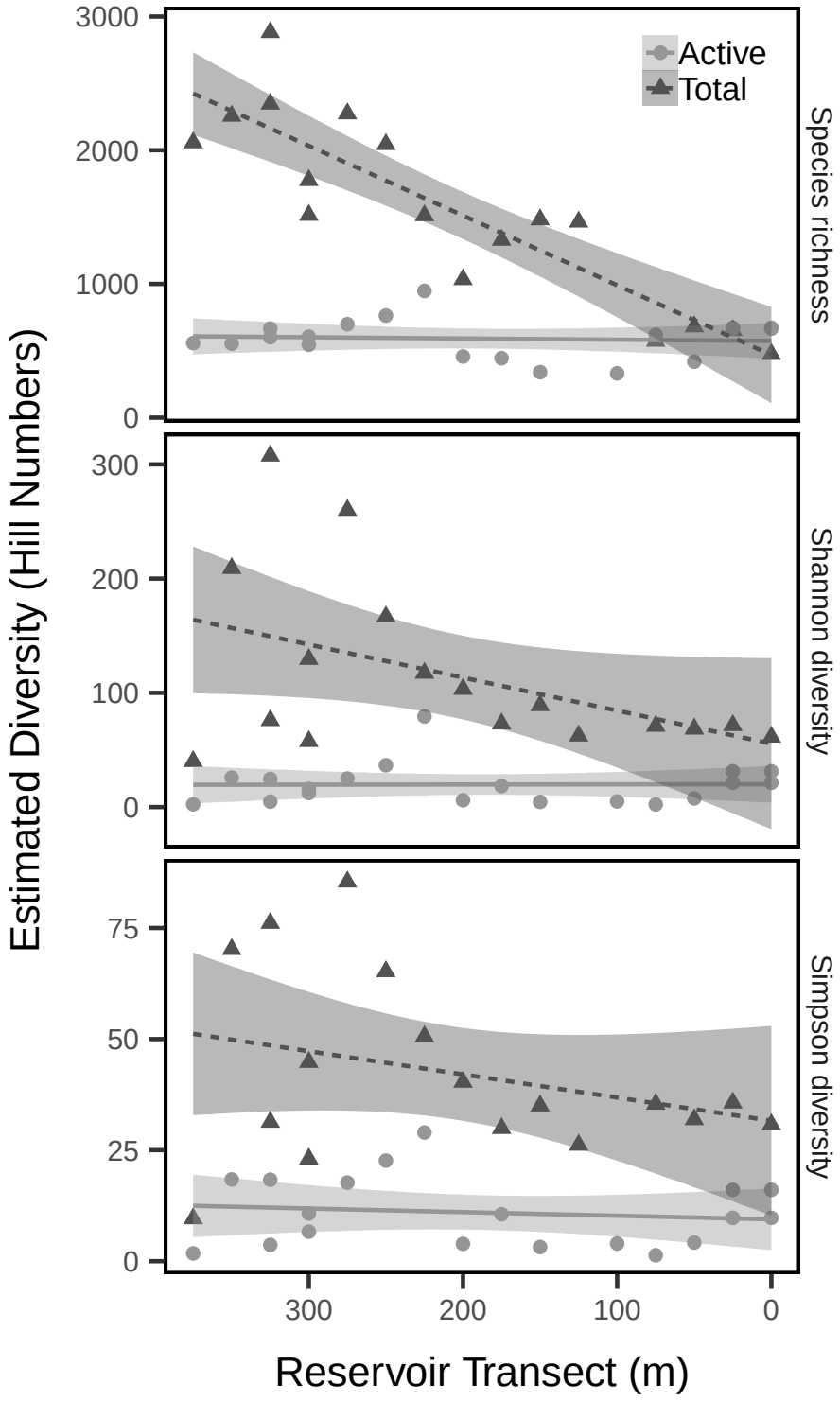
Table 1: Table continues below

Diversity	Term	Estimate	Std. Error	Statistic
Species richness	distance	5.21	0.54	9.66
Species richness	moleculeRNA	105	166	0.632
Species richness	distance:moleculeRNA	-5.12	0.716	-7.15
Shannon diversity	distance	0.289	0.105	2.75
Shannon diversity	moleculeRNA	-35.5	32.3	-1.1
Shannon diversity	distance:moleculeRNA	-0.291	0.139	-2.08
Simpson diversity	distance	0.0522	0.0313	1.66
Simpson diversity	moleculeRNA	-22.2	9.65	-2.3
Simpson diversity	distance:moleculeRNA	-0.0441	0.0416	-1.06

p-value
5.22e-11
0.532
4.07e-08
0.00965
0.28
0.0453
0.106
0.0282
0.297

```
hill.estim %>% filter(type == "water") %>%
  mutate(molecule = ifelse(molecule == "DNA", "Total", "Active")) %>%
  ggplot(aes(x = distance, y = Estimator,
             ymin = LCL, ymax = UCL,
             color = molecule, fill = molecule, shape = molecule)) +
  geom_point(size =3) +
  # geom_errorbar(size = .5, aes(ymin = Estimator - s.e., ymax = Estimator + s.e.),
  #                width = 10, alpha = 0.5) +
  geom_smooth(method = "lm", aes(linetype = molecule)) +
  labs(x = "Reservoir Transect (m)",
       y = "Estimated Diversity (Hill Numbers)") +
  scale_color_manual(values = my.cols) +
  scale_fill_manual(values = my.cols) +
  theme(legend.position = c(.85,.95)) +
```

```
  scale_x_reverse() +
  facet_grid(Diversity ~ ., scales = "free")
```



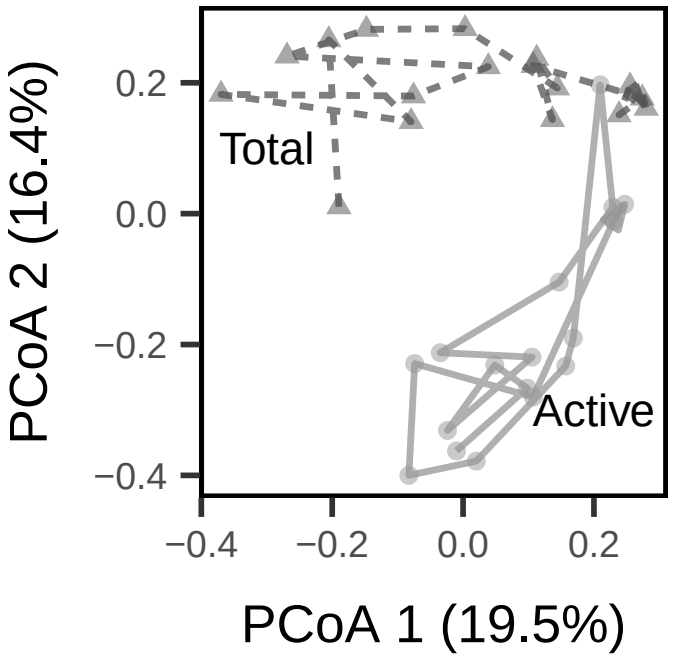
So, from the basis of these results, we can make the following conclusions. First, we note that diversity in the total community decays from the stream inlet to the dam of the reservoir. That is, all the lines have a negative slope. However, we do not see this decay in the metabolically active community. Second, we note that the metabolically actively community has much lower diversity than the total community near the soils,

but this difference decreases toward the dam. Last, because we quantified diversity across three orders of Hill numbers (q = 0, 1, and 2), we can also say something about the relative importance of rare versus common taxa along the reservoir transect. We see the the significance of the distance-by-molecule interaction term decrease as rare taxa are downweighted in favor of common taxa. This suggests that the differences between the active and total communities along the transect is driven primarily by rare taxa. However, the general trend of higher Simpson diversity across the whole transect suggests that low-activity, but relatively common, taxa are maintained in the reservoir.

How does community structure change along the gradient?

First, we'll just get an overview of how the communities look along the aquatic transect.

```
ul.pcoa <- cmdscale(vegdist(OTUsREL.log, method="bray"), 2, eig = T, add = T)
explainvars <- round(eigenvals(ul.pcoa)[c(1,2)]/sum(eigenvals(ul.pcoa)),3) *100
water.pcvals <- data.frame(scores(ul.pcoa)) %>%
  rownames_to_column("name") %>%
  left_join(rownames_to_column(design, "name")) %>%
  arrange(desc(distance)) %>% filter(type == "water")
pc_dists <- tibble(
  DNA_dim1 = subset(water.pcvals, molecule == "DNA")$Dim1,
  DNA_dim2 = subset(water.pcvals, molecule == "DNA")$Dim2,
  RNA_dim1 = subset(water.pcvals, molecule == "RNA")$Dim1,
  RNA_dim2 = subset(water.pcvals, molecule == "RNA")$Dim2)
data.frame(scores(ul.pcoa)) %>%
  rownames_to_column("name") %>%
  left_join(rownames_to_column(design, "name")) %>%
  arrange(desc(distance)) %>% filter(type == "water") %>%
  mutate(molecule = ifelse(molecule == "DNA", "Total", "Active")) %>%
  ggplot(aes(x = Dim1, y = Dim2)) +
  geom_point(size = 3, alpha = 0.5, aes(color = molecule, shape = molecule)) +
  geom_path(size = 1.2, alpha = 0.75, arrow = arrow(angle = 20,
                          length = unit(0.35, "cm"),
                          type = "closed"), aes(color = molecule, linetype = molecule)) +
  scale_color_manual("Community Subset", values = my.cols) +
  geom_segment(data = pc_dists,
               aes(x = DNA_dim1, y = DNA_dim2,
                   xend = RNA_dim1, yend = RNA_dim2),
               alpha = 0) +
  coord_fixed() +
  labs(x = paste0("PCoA 1 (", explainvars[1],"%)"),
       y = paste0("PCoA 2 (", explainvars[2],"%)")) +
  theme(legend.position = "none") +
  annotate(geom = "text", x = .2, y = -.3, label = "Active", size = 6) +
  annotate(geom = "text", x = -.3, y = .1, label = "Total", size = 6) +
  ggsave("figures/active-tot-pcoa-trajectories.pdf", bg = "white", width = 8, height = 8) +
  ggsave("figures/active-tot-pcoa-trajectories.png", width = 8, height = 8)
```




```
# animation
# traj.animate <- data.frame(scores(ul.pcoa)) %>%
#   rownames_to_column("name") %>%
#   left_join(rownames_to_column(design, "name")) %>%
#   arrange(desc(distance)) %>% filter(type == "water", station != "UL17", station != "UL18") %>%
#   mutate(molecule = ifelse(molecule == "DNA", "Total", "Active")) %>%
#   mutate(transect = desc(distance)) %>%
#   ggplot(aes(x = Dim1, y = Dim2, frame = transect, cumulative = TRUE)) +
#   geom_point(alpha = 0.75, size = 10, aes(color = molecule)) +
#   geom_path(alpha = 1, size = 1, arrow = arrow(angle = 20,
#                            length = unit(0.35, "cm"),
#                            type = "closed"), aes(color = molecule)) +
#   scale_color_manual(values = my.cols) +
#   theme(legend.title = element_blank()) +
#   coord_fixed() +
#   labs(x = paste0("PCoA 1 (", explainvars[1],"%)"),
#        y = paste0("PCoA 2 (", explainvars[2],"%)"))
#
# gganimate(traj.animate, filename = "../figures/trajectory-animation.mp4", ani.width = 800, ani.height
```

So, it appears that there is convergence in community structure along the path from stream inlet to the dam. This could reflect a loss of soil-derived taxa in the aquatic samples. To test this, we'll look at β-diversity along the gradient with respect to the soil samples. If we see a decay in similarity to soils, this suggests soil taxa are having a comparatively lower influence with distance from the inlet.

Similarity To Terrestrial Habitat Across Gradient (Terrestrial Influence)

```
# Similarity to Soil Sample
UL.bray      <- 1-as.matrix(vegdist(OTUsREL.log, method="bray"))
UL.bray.lake <- UL.bray[-c(1:3), 1:3]
bray.mean    <- round(apply(UL.bray.lake, 1, mean), 3)
bray.se      <- round(apply(UL.bray.lake, 1, se), 3)
```

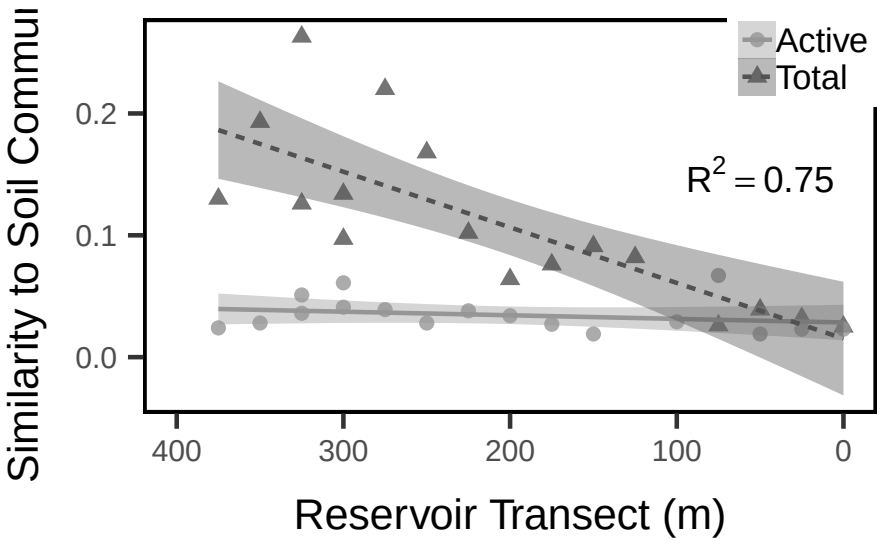
```
UL.sim       <- cbind(design[-c(1:3), ], bray.mean, bray.se)

# Calculate Linear Model
model.terr <- lm(bray.mean ~ distance * molecule, data = UL.sim)
pander(model.terr)
```

Table 3: Fitting linear model: bray.mean ~ distance * molecule

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.01524	0.01623	0.9392	0.3551
distance	0.0004564	6.828e-05	6.684	2.097e-07
moleculeRNA	0.01321	0.02281	0.579	0.5669
distance:moleculeRNA	-0.0004269	9.608e-05	-4.443	0.0001117

```
# # Calculate Confidance Intervals of Model
# newdata.terr <- data.frame(cbind(UL.sim$molecule, UL.sim$distance))
# conf95.terr <- predict(model.terr, newdata.terr, interval="confidence")
#
# # Dummy Variables Regression Model ("Terrestrial Influence")
# D2 <- (UL.sim$molecule == "RNA")*1
# fit.Fig.3b <- lm(UL.sim$bray.mean ~ UL.sim$distance + D2 + UL.sim$distance*D2)
# D2.R2 <- round(summary(fit.Fig.3b)$r.squared, 2)
# summary(fit.Fig.3b)
#
#
# DNA.int.3b <- fit.Fig.3b$coefficients[1]
# DNA.slp.3b <- fit.Fig.3b$coefficients[2]
# RNA.int.3b <- DNA.int.3b + fit.Fig.3b$coefficients[3]
# RNA.slp.3b <- DNA.slp.3b + fit.Fig.3b$coefficients[4]
```

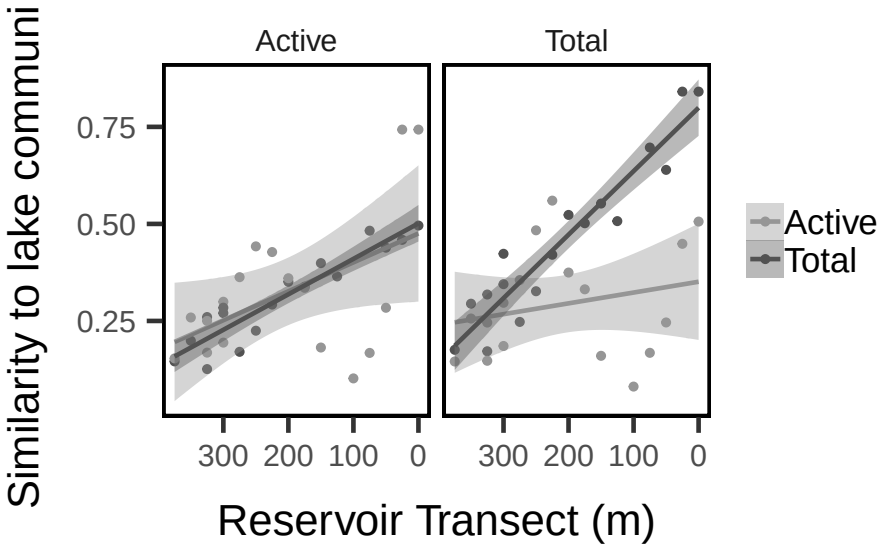
```
UL.sim %>%
  mutate(molecule = ifelse(UL.sim$molecule == "DNA", "Total", "Active")) %>%
  ggplot(aes(x = distance, y = bray.mean,
             color = molecule, fill = molecule, shape = molecule)) +
  geom_point(alpha = 0.8, size = 3, show.legend = T) +
  geom_smooth(method = "lm", show.legend = T, aes(linetype = molecule)) +
  labs(y = "Similarity to Soil Community",
       x = "Reservoir Transect (m)") +
  scale_color_manual(values = my.cols) +
  scale_fill_manual(values = my.cols) +
  theme(legend.position = c(0.9, 0.9)) +
  scale_x_reverse(limits = c(400,0)) +
  annotate(geom = "text", x = 50, y = 0.15, size = 6,
           label = paste0("R^2== ",round(summary(model.terr)$r.squared, 2)), parse = T)
```



What about within-lake β-diversity?

An alternative reason for convergence of aquatic communities along the reservoir and decay in similarity to soils is that niche selection is acting on aquatic community structure and driving shifts in composition along the gradient. To test this idea, we look at the similarity of the active and total communities to the samples taken at the dam of the reservoir, which we expect to have a low influence of soil taxa.

```
# Similarity to Lake Samples 1 and 2
UL.bray2      <- 1 - as.matrix(vegdist(OTUsREL.log, method="bray"))
UL.bray.lake2 <- UL.bray[-c(1:3), 4:7]
UL.sim2       <- cbind(design[-c(1:3), ],
                       "DNA" = apply(UL.bray.lake2[,c(1,3)], 1, mean),
                       "RNA" = apply(UL.bray.lake2[,c(2,4)], 1, mean))

UL.sim2 %>% gather(DNA, RNA, key = "comparison", value = "similarity") %>%
  mutate(molecule = ifelse(molecule == "DNA", "Total", "Active"),
         comparison = ifelse(comparison == "DNA", "Total", "Active")) %>%
  ggplot(aes(x = distance, y = similarity, color = molecule, fill = molecule)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(y = "Similarity to lake community",
       x = "Reservoir Transect (m)") +
  scale_x_reverse() +
  scale_color_manual("Community Subset", values = my.cols) +
  scale_fill_manual("Community Subset", values = my.cols) +
  facet_wrap(~ comparison)
```




```
# Calculate Linear Model
model.lake1 <- lm(UL.sim2$DNA ~ UL.sim2$distance * UL.sim2$molecule)
model.lake2 <- lm(UL.sim2$RNA ~ UL.sim2$distance * UL.sim2$molecule)
summary(model.lake1)
```

```
##
## Call:
## lm(formula = UL.sim2$DNA ~ UL.sim2$distance * UL.sim2$molecule)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.242239 -0.086327  0.000567  0.063327  0.271642
##
## Coefficients:
##                                       Estimate Std. Error t value
## (Intercept)                           0.7994111  0.0558603  14.311
## UL.sim2$distance                     -0.0016356  0.0002350  -6.960
## UL.sim2$moleculeRNA                  -0.4484387  0.0784869  -5.714
## UL.sim2$distance:UL.sim2$moleculeRNA  0.0013571  0.0003307   4.104
##                                      Pr(>|t|)
## (Intercept)                          6.08e-15 ***
## UL.sim2$distance                     9.88e-08 ***
## UL.sim2$moleculeRNA                  3.11e-06 ***
## UL.sim2$distance:UL.sim2$moleculeRNA 0.000286 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1126 on 30 degrees of freedom
## Multiple R-squared:  0.6954, Adjusted R-squared:  0.6649
## F-statistic: 22.83 on 3 and 30 DF,  p-value: 6.829e-08
```

```
summary(model.lake2)
```

```
##
## Call:
## lm(formula = UL.sim2$RNA ~ UL.sim2$distance * UL.sim2$molecule)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.298879 -0.046626 -0.003357  0.046799  0.286274
##
## Coefficients:
##                                     Estimate Std. Error t value
## (Intercept)                          0.5013467  0.0610169   8.217
## UL.sim2$distance                    -0.0009140  0.0002567  -3.561
## UL.sim2$moleculeRNA                 -0.0259821  0.0857323  -0.303
## UL.sim2$distance:UL.sim2$moleculeRNA  0.0001688  0.0003612   0.467
##                                     Pr(>|t|)
## (Intercept)                         3.59e-09 ***
## UL.sim2$distance                     0.00126 **
## UL.sim2$moleculeRNA                  0.76393
## UL.sim2$distance:UL.sim2$moleculeRNA  0.64353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.123 on 30 degrees of freedom
## Multiple R-squared:  0.4156, Adjusted R-squared:  0.3572
## F-statistic: 7.112 on 3 and 30 DF,  p-value: 0.0009525
```

```
# # Calculate Confidance Intervals of Model
# newdata.lake <- data.frame(cbind(UL.sim2$molecule, UL.sim2$distance))
# conf95.lake <- predict(model.lake1, newdata.lake, interval="confidence")
#
# # Dummy Variables Regression Model ("Lake Influence")
# D3 <- (UL.sim2$molecule == "RNA")*1
# fit.Fig.3c <- lm(UL.sim2$DNA ~ UL.sim2$distance + D3 + UL.sim2$distance*D3)
# # summary(fit.Fig.3c)
#
# DNA.int.3c <- fit.Fig.3c$coefficients[1]
# DNA.slp.3c <- fit.Fig.3c$coefficients[2]
# RNA.int.3c <- DNA.int.3c + fit.Fig.3c$coefficients[3]
# RNA.slp.3c <- DNA.slp.3c + fit.Fig.3c$coefficients[4]
```

Identifying the Soil Bacteria

```
soil.only <- OTUs[, which(colSums(OTUs[-c(1:3),]) == 0)] # what is present only in soil
lake.n.soil <- OTUs[, setdiff(colnames(OTUs),colnames(soil.only))] # what is present across all samples


# separate lake and soil samples
lake.total <- OTUs[which(design$molecule == "DNA", design$type == "water"),]
soil.total <- OTUs[which(design$molecule == "DNA", design$type == "soil"),]

# which otus are present in both lake and soil samples
lake.and.soil.total <- OTUs[which(design$molecule == "DNA", design$type == "water"),
                            which(colSums(lake.total) > 0 & colSums(soil.total) > 0)]

# isolate just the dna and rna lake communities
w.dna <- OTUs[which(design$molecule == "DNA" & design$type == "water"), ]
```

```
w.rna <- OTUs[which(design$molecule == "RNA" & design$type == "water"), ]

# pull out the lake rna counds for otus found in lake and soil
lake.and.soil.act <- w.rna[,colnames(lake.and.soil.total)]

# of these lake and soil taxa, which are never active?
nvr.act <- which(colSums(lake.and.soil.act) == 0)

# pull out their dna abundances and calculate richness
terr.lake <- w.dna[ , c(names(nvr.act))]
terr.rich <- rowSums((terr.lake > 0) * 1)
terr.REL <- rowSums(terr.lake) / rowSums(w.dna)
design.dna <- design[which(design$molecule == "DNA" & design$type == "water"), ]
terr.rich.log <- log10(terr.rich)
terr.REL.log <- log10(terr.REL)

terr.mod1 <- lm(terr.rich.log ~ design.dna$distance)
summary(terr.mod1)
```

```
##
## Call:
## lm(formula = terr.rich.log ~ design.dna$distance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23774 -0.13355  0.05749  0.10655  0.22647
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.0158897  0.0771478   26.13 6.36e-14 ***
## design.dna$distance 0.0031610  0.0003245    9.74 7.06e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1555 on 15 degrees of freedom
## Multiple R-squared:  0.8635, Adjusted R-squared:  0.8544
## F-statistic: 94.87 on 1 and 15 DF,  p-value: 7.059e-08
```

```
T1.R2 <- round(summary(terr.mod1)$r.squared, 2)
T1.int <- terr.mod1$coefficients[1]
T1.slp <- terr.mod1$coefficients[2]
pander(terr.mod1)
```

Table 4: Fitting linear model: terr.rich.log ~ design.dna$distance

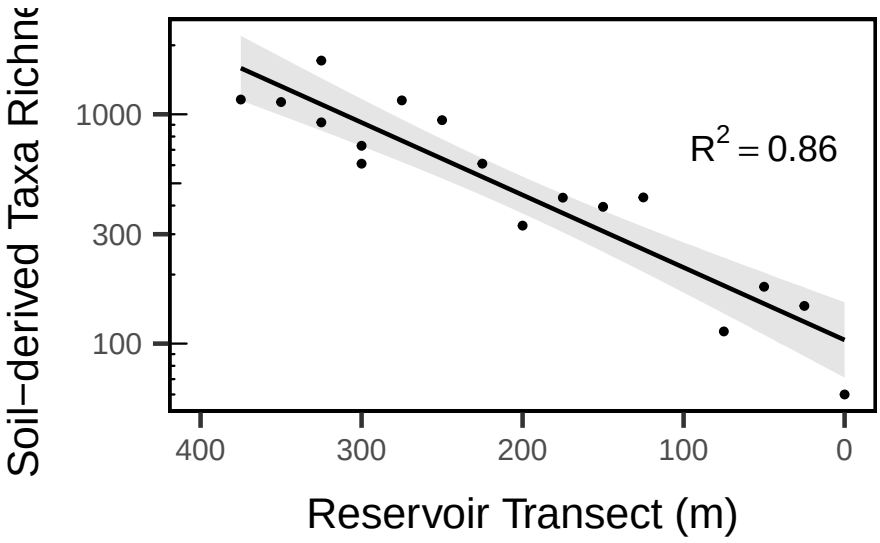
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.016	0.07715	26.13	6.362e-14
design.dna$distance	0.003161	0.0003245	9.74	7.059e-08

```
# terr.mod2 <- lm(terr.REL.log ~ design.dna$distance)
# summary(terr.mod2)
# T2.R2 <- round(summary(terr.mod2)$r.squared, 2)
# T2.int <- terr.mod2$coefficients[1]
```

```
# T2.slp <- terr.mod2$coefficients[2]
```

Figure 4: Transient decay

```
tibble(transient_rich = terr.rich, distance = design.dna$distance) %>%
  ggplot(aes(x = distance, y = transient_rich)) +
  geom_smooth(method = "lm", color = "black", fill = "grey") +
  geom_point(alpha = 1, color = "black") +
  scale_x_reverse(limits = c(400,0)) +
  scale_y_log10() +
  annotation_logticks(sides = "l") +
  labs(x = "Reservoir Transect (m)",
       y = "Soil-derived Taxa Richness") +
  annotate("text", x = 50, y = 750, size = 6, label = paste0("R^2== ",T1.R2), parse = T)
```



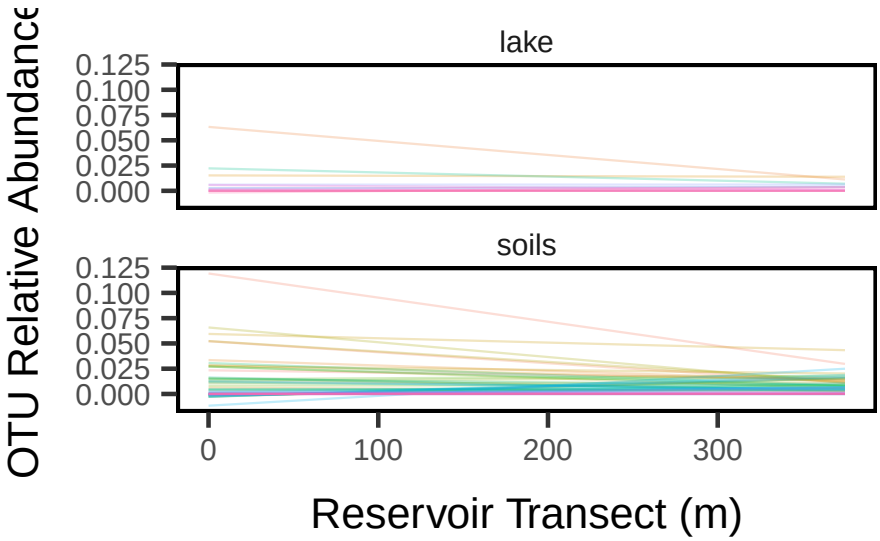
Define Core Lake Taxa

```
# identify otus in soil samples and lake samples
in.soil <- OTUs[, which(colSums(OTUs[c(1:3),]) > 0 )]
in.lake <- OTUs[, which(colSums(OTUs[-c(1:3),]) > 0)]

# isolate just the rna water samples and convert to presence-absence
in.lake.rna <- in.lake[which(design$molecule == "RNA" & design$type == "water"), ]
in.lake.rna.pa <- (in.lake.rna > 0) * 1

# define the 'core' taxa as otus present in 90% of samples
in.lake.core <- w.dna[, which((colSums(in.lake.rna.pa) / nrow(in.lake.rna.pa)) >= 0.8)]

# of the core, how many are also in the soil samples?
in.lake.core.from.soils <- in.lake.core[, intersect(colnames(in.lake.core), colnames(in.soil))]

# of the core which are not in the soil samples
```

```
in.lake.core.not.soils <- in.lake.core[, setdiff(colnames(in.lake.core), colnames(in.soil))]


in.lake.core.from.soils.REL <- in.lake.core.from.soils / rowSums(w.dna)

in.soil.to.plot <- as.data.frame(in.lake.core.from.soils.REL) %>%
  rownames_to_column("sample_ID") %>%
  gather(otu_id, rel_abundance, -sample_ID) %>%
  left_join(rownames_to_column(design.dna, "sample_ID")) %>%
  add_column(found = "soils")

in.lake.core.not.soils.REL <- in.lake.core.not.soils / rowSums(w.dna)

in.lake.to.plot <- as.data.frame(in.lake.core.not.soils.REL) %>%
  rownames_to_column("sample_ID") %>%
  gather(otu_id, rel_abundance, -sample_ID) %>%
  left_join(rownames_to_column(design.dna, "sample_ID")) %>%
  add_column(found = "lake")
```

Now, lets plot the abundances of the OTUs across the reservoir and split them up into whether they were recovered in soils or not.

```
bind_rows(in.soil.to.plot, in.lake.to.plot) %>%
  ggplot(aes(x = distance, y = rel_abundance, color = otu_id)) +
  labs(x = "Reservoir Transect (m)",
       y = "OTU Relative Abundance") +
  geom_line(alpha = 0.25, stat = "smooth", method = "lm", se = F, show.legend = F) +
  facet_wrap(~ found, ncol = 1)
```




```
data.frame(mean_relabund = colMeans(in.lake.core.from.soils.REL)) %>%
  rownames_to_column(var = "OTU") %>% left_join(OTU.tax) %>%
  mutate(Taxon = paste(Phylum, Class, Order)) %>%
  arrange(desc(mean_relabund)) %>%
  ggplot() +
  geom_bar(aes(x = Taxon, y = mean_relabund), stat = "identity") +
  coord_flip()
```

```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```


```
soil.core.mods <- apply(in.lake.core.from.soils.REL, MARGIN = 2,
    FUN = function(x) summary(lm(x ~ design.dna$distance))$coefficients[2,c(1,4)])
rownames(soil.core.mods) <- c("slope", "pval")
soil.core.decresing <- as.data.frame(t(soil.core.mods)) %>%
  rownames_to_column("OTU") %>%
  filter(pval < 0.05 & slope > 0) %>%   # rel abund decreases toward dam
  left_join(OTU.tax)
```

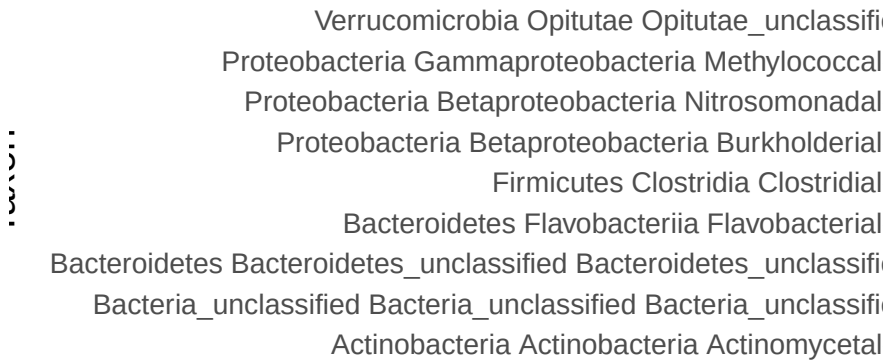
```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```

```
soil.core.increasing <- as.data.frame(t(soil.core.mods)) %>%
  rownames_to_column("OTU") %>%
  filter(pval < 0.05 & slope < 0) %>%   # rel abund increases toward dam
  left_join(OTU.tax)
```

```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```

```
data.frame(mean_relabund = colMeans(in.lake.core.not.soils.REL)) %>%
  rownames_to_column(var = "OTU") %>% left_join(OTU.tax) %>%
  mutate(Taxon = paste(Phylum, Class, Order)) %>%
  arrange(desc(mean_relabund)) %>%
  ggplot() +
  geom_bar(aes(x = Taxon, y = mean_relabund), stat = "identity") +
  coord_flip()
```

```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```



mear

```
nonsoil.core.mods <- apply(in.lake.core.not.soils.REL, MARGIN = 2, FUN = function(x) summary(lm(x ~ des
rownames(nonsoil.core.mods) <- c("slope", "pval")
nonsoil.core.decreasing <- as.data.frame(t(nonsoil.core.mods)) %>%
  rownames_to_column("OTU") %>%
  filter(pval < 0.05 & slope > 0) %>%   # rel abund decreases toward dam
  left_join(OTU.tax)
```

```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```

```
nonsoil.core.increasing <- as.data.frame(t(nonsoil.core.mods)) %>%
  rownames_to_column("OTU") %>%
  filter(pval < 0.05 & slope < 0) %>%   # rel abund increases toward dam
  left_join(OTU.tax)
```

```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```

```
as.data.frame(OTUsREL[,nonsoil.core.increasing$OTU]) %>% rownames_to_column("sampleID") %>%
  left_join(rownames_to_column(design, "sampleID")) %>%
  gather(OTU, rel_abund, -station, -molecule, -type, -distance, -sampleID) %>%
  filter(molecule == "DNA") %>% left_join(OTU.tax) %>%
  mutate(taxon = paste(Class, Order, Family)) %>%
  ggplot(aes(x = distance, y = rel_abund, color = taxon)) +
  geom_point() +
  geom_smooth(method = "lm") +
  scale_x_reverse()
```

```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```

```
## Warning: Removed 15 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 15 rows containing missing values (geom_point).
```

:tinobacteria Actinomycetales Actinomycetales_unclassified
ıcteroidetes_unclassified Bacteroidetes_unclassified Bacteroi

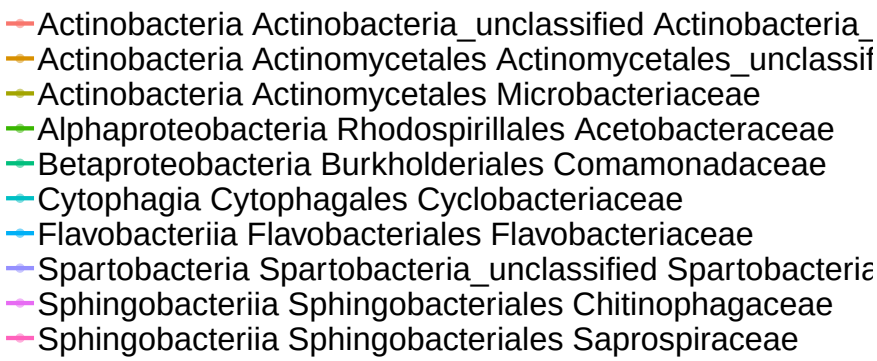
```
as.data.frame(OTUsREL[,soil.core.increasing$OTU]) %>% rownames_to_column("sampleID") %>%
  left_join(rownames_to_column(design, "sampleID")) %>%
  gather(OTU, rel_abund, -station, -molecule, -type, -distance, -sampleID) %>%
  filter(molecule == "DNA") %>% left_join(OTU.tax) %>%
```

```
  mutate(taxon = paste(Class, Order, Family)) %>%
  ggplot(aes(x = distance, y = rel_abund, color = taxon)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_x_reverse()
```

```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```

```
## Warning: Removed 42 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 42 rows containing missing values (geom_point).
```



ce

```
# how much do the different core components contribute to total abundances
in.lake.core.soil.REL <- rowSums(in.lake.core.from.soils) / rowSums(w.dna)
in.lake.core.water.REL <- rowSums(in.lake.core.not.soils) / rowSums(w.dna)
```

Taxonomic Analysis

```
# Taxa comprising total lake 'core', those from soils, and those not from soils
core.taxa <- OTU.tax[OTU.tax$OTU %in% colnames(in.lake.core),]

core.soil.taxa <- OTU.tax[OTU.tax$OTU %in% colnames(in.lake.core.from.soils),]
core.water.taxa <- OTU.tax[OTU.tax$OTU %in% colnames(in.lake.core.not.soils),]

# Get relative abundances for each of the core taxa
core.soil.taxa.DNA.REL <- OTUsREL[which(design$molecule == "DNA" & design$type == "water"),
                                  as.numeric(rownames(core.soil.taxa))]
core.water.taxa.DNA.REL <- OTUsREL[which(design$molecule == "DNA" & design$type == "water"),
                                   as.numeric(rownames(core.water.taxa))]
core.soil.taxa.RNA.REL <- OTUsREL[which(design$molecule == "RNA" & design$type == "water"),
                                  as.numeric(rownames(core.soil.taxa))]
core.water.taxa.RNA.REL <- OTUsREL[which(design$molecule == "RNA" & design$type == "water"),
                                   as.numeric(rownames(core.water.taxa))]

core.soil.taxa.DNA.REL.max <- as.matrix(apply(core.soil.taxa.DNA.REL, 2, max))
core.soil.taxa.RNA.REL.max <- as.matrix(apply(core.soil.taxa.RNA.REL, 2, max))
core.water.taxa.DNA.REL.max <- as.matrix(apply(core.water.taxa.DNA.REL, 2, max))
```

```r
core.water.taxa.RNA.REL.max <- as.matrix(apply(core.water.taxa.RNA.REL, 2, max))

core.soil.taxa.DNA.REL.min <- as.matrix(apply(core.soil.taxa.DNA.REL, 2, min))
core.soil.taxa.RNA.REL.min <- as.matrix(apply(core.soil.taxa.RNA.REL, 2, min))
core.water.taxa.DNA.REL.min <- as.matrix(apply(core.water.taxa.DNA.REL, 2, min))
core.water.taxa.RNA.REL.min <- as.matrix(apply(core.water.taxa.RNA.REL, 2, min))

core.soil.taxa.DNA.REL.mean <- as.matrix(apply(core.soil.taxa.DNA.REL, 2, mean))
core.soil.taxa.RNA.REL.mean <- as.matrix(apply(core.soil.taxa.RNA.REL, 2, mean))
core.water.taxa.DNA.REL.mean <- as.matrix(apply(core.water.taxa.DNA.REL, 2, mean))
core.water.taxa.RNA.REL.mean <- as.matrix(apply(core.water.taxa.RNA.REL, 2, mean))

core.soil.taxa.soil.max <- as.matrix(apply(OTUsREL[which(design$type == "soil"), rownames(core.soil.taxa

core.soil.taxa.DNA.REL.bounds <- cbind(core.soil.taxa.DNA.REL.min, core.soil.taxa.DNA.REL.max,
                                       core.soil.taxa.RNA.REL.min, core.soil.taxa.RNA.REL.max,
                                       core.soil.taxa.DNA.REL.mean, core.soil.taxa.RNA.REL.mean,
                                       core.soil.taxa.soil.max)

colnames(core.soil.taxa.DNA.REL.bounds) <- c("DNA.min", "DNA.max", "RNA.min", "RNA.max", "DNA.mean", "R




core.water.taxa.DNA.REL.bounds <- cbind(core.water.taxa.DNA.REL.min, core.water.taxa.DNA.REL.max,
                                       core.water.taxa.RNA.REL.min, core.water.taxa.RNA.REL.max,
                                       core.water.taxa.DNA.REL.mean, core.water.taxa.RNA.REL.mean)
colnames(core.water.taxa.DNA.REL.bounds) <- c("DNA.min", "DNA.max", "RNA.min", "RNA.max", "DNA.mean", "

# core.soil and core.water are summaries of lake core
core.soil <- as.data.frame(cbind(core.soil.taxa$Family, core.soil.taxa$Genus,
                                 signif(core.soil.taxa.DNA.REL.bounds[,c(1:4, 7)], digits = 3)))
colnames(core.soil) <- c("Family", "Genus", "DNA.min", "DNA.max", "RNA.min", "RNA.max", "Soil.max")
core.water <- as.data.frame(cbind(core.water.taxa$Family, core.water.taxa$Genus,
                                  signif(core.water.taxa.DNA.REL.bounds[,1:4], digits = 3)))
colnames(core.water) <- c("Family", "Genus", "DNA.min", "DNA.max", "RNA.min", "RNA.max")

# Core Soil LaTeX Table
addtorow <- list()
addtorow$pos <- list(0, 0)
addtorow$command <- c("& \\multicolumn{1}{c}{Class} & \\multicolumn{1}{c}{Order} &
                      \\multicolumn{2}{c}{DNA} & \\multicolumn{2}{c}{RNA} \\\\\n",
                      "& &  & min & max & min & max \\\\\n")
core.soil.tab <- xtable(core.soil)
align(core.soil.tab) <- "crrrrrrr"
print(core.soil.tab, add.to.row = addtorow, include.colnames = FALSE,
      type= "latex", file="tables/table1.tex")
print(core.soil.tab, add.to.row = addtorow, include.colnames = FALSE, comment = FALSE)
```

```r
core.water.tab <- xtable(core.water)
align(core.water.tab) <- "crrrrrr"
print(core.water.tab, add.to.row = addtorow, include.colnames = FALSE,
      type= "latex", file="tables/table2.tex")
print(core.water.tab, add.to.row = addtorow, include.colnames = FALSE, comment = FALSE)
```

Comparisons of relabunds

Now, lets see which taxa increase or decrease substantially along the gradient. I calculated the fold change in relative abundance of all these taxa along the gradient relative to their max abundance in soils. Thus, the OTUs that are most abundant near the soils will have a declining slope toward the dam. The OTUs that are perhaps seeded from the soils into the lake will have an increasing slope toward the dam.

```
high.activity.soil.core <- as.data.frame(core.soil.taxa.DNA.REL.bounds) %>%
  rownames_to_column("OTU") %>%
  filter(RNA.max > 0) %>% arrange(desc(RNA.max)) %>%
  left_join(OTU.tax)
```

```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```

```
high.activity.water.core <- as.data.frame(core.water.taxa.DNA.REL.bounds) %>%
  rownames_to_column("OTU") %>%
  filter(RNA.max > 0) %>% arrange(desc(RNA.max)) %>%
  left_join(OTU.tax)
```

```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```

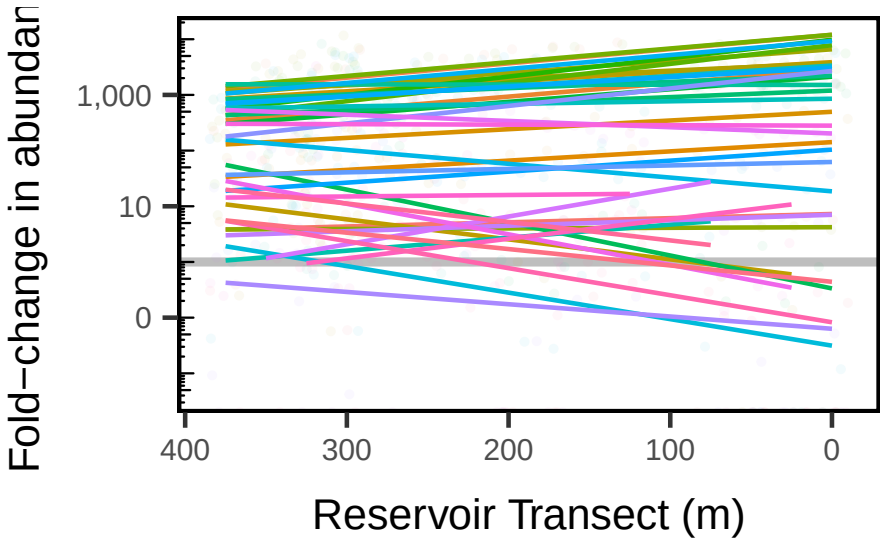
```
mean.soil.abunds.soil.core <- OTUsREL[which(design$type == "soil"), high.activity.soil.core$OTU] %>%
  colMeans %>% data.frame(mean_soil_relabund = .) %>%
  rownames_to_column("OTU") %>% arrange(desc(mean_soil_relabund))
max.soil.abunds.soil.core <- OTUsREL[which(design$type == "soil"), high.activity.soil.core$OTU] %>%
  apply(X = ., MARGIN = 2, max) %>% data.frame(max_soil_relabund = .) %>%
  rownames_to_column("OTU") %>% arrange(desc(max_soil_relabund))

mean.soil.abunds.water.core <- OTUsREL[which(design$type == "soil"), high.activity.water.core$OTU] %>%
  colMeans %>% data.frame(mean_soil_relabund = .) %>%
  rownames_to_column("OTU") %>% arrange(desc(mean_soil_relabund))
max.soil.abunds.water.core <- OTUsREL[which(design$type == "soil"), high.activity.water.core$OTU] %>%
  apply(X = ., MARGIN = 2, max) %>% data.frame(max_soil_relabund = .) %>%
  rownames_to_column("OTU") %>% arrange(desc(max_soil_relabund))


soil.vs.lake.abunds <- high.activity.soil.core %>%
  left_join(mean.soil.abunds.soil.core) %>% left_join(max.soil.abunds.soil.core) %>%
  mutate(soil_is_source = ifelse(max_soil_relabund > 1e-3 & RNA.max > 1e-3, T, F)) %>%
  mutate(Taxon = ifelse(Genus == "unclassified", paste(Family, "sp."), Genus))

combined.relabunds <- max.soil.abunds.soil.core %>%
  left_join(rownames_to_column(as.data.frame(t(in.lake.core.from.soils.REL)), "OTU"))
rownames(combined.relabunds) <- combined.relabunds$OTU
combined.relabunds <- combined.relabunds[,-1]

otus.fold.change <- na.omit(combined.relabunds / combined.relabunds$max_soil_relabund) # Calculate fold

fold_change_summary <- otus.fold.change %>% rownames_to_column("OTU") %>%
  select(-max_soil_relabund) %>%
  gather("sample", "fold_change", -OTU) %>%
```

```
  left_join(select(rownames_to_column(design.dna, "sample"), -station, -molecule, -type)) %>%
  group_by(OTU) %>%
  summarize(max_change = max(fold_change), min_change = min(fold_change))

otus.fold.change %>% rownames_to_column("OTU") %>%
  select(-max_soil_relabund) %>%
  gather("sample", "fold_change", -OTU) %>%
  left_join(select(rownames_to_column(design.dna, "sample"), -station, -molecule, -type)) %>%
  ggplot(aes(x = distance, y = fold_change, color = OTU)) +
  geom_hline(aes(yintercept = 1), color = "gray50", alpha = 0.5, size = 2) +
  geom_jitter(alpha = 0.05) +
  geom_smooth(alpha = 0.5, method = "lm", se = F) +
  scale_y_log10(labels = scales::comma) +
  scale_x_reverse() +
  annotation_logticks(long = unit(.1, "in"), sides = "l") +
  theme(legend.position = "none") +
  labs(x = "Reservoir Transect (m)", y = "Fold-change in abundance")
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 42 rows containing non-finite values (stat_smooth).
```




```
# otus.fold.change %>% rownames_to_column("OTU") %>%
#   select(-max_soil_relabund) %>%
#   gather("sample", "fold_change", -OTU) %>%
#   left_join(select(rownames_to_column(design.dna, "sample"), -station, -molecule, -type))

foldchanges <- t(otus.fold.change)[-1,]
foldchangelms <- apply(foldchanges, MARGIN = 2,
    FUN = function(x) summary(lm(x ~ design.dna$distance))$coefficients[c(1,2,8)])
rownames(foldchangelms) <- c("intercept", "slope", "pval")

soil.core.decresing <- as.data.frame(t(foldchangelms)) %>%
  rownames_to_column("OTU") %>%
  filter( slope > 0) %>%   # rel abund decreases toward dam
```

```
  left_join(OTU.tax) %>% select(-intercept, -slope, -pval, everything()) %>%
  arrange(desc(slope))
```

```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```

```
soil.core.increasing <- as.data.frame(t(foldchangelms)) %>%
  rownames_to_column("OTU") %>%
  filter( slope < 0) %>%   # rel abund increases toward dam
  left_join(OTU.tax) %>% select(-intercept, -slope, -pval, everything()) %>%
  arrange((slope))
```

```
## Warning: Column `OTU` joining character vector and factor, coercing into
## character vector
```

```
soil.decrease.tab <- soil.core.decresing %>% select(-OTU, -Domain) %>%  flextable()
soil.increase.tab <- soil.core.increasing %>% select(-OTU, -Domain) %>% flextable()

read_docx() %>%
  body_end_section_continuous() %>%
  body_add_par("Increasing away from stream inlet", style = "heading 2") %>%
  body_add_flextable(soil.increase.tab) %>%
  body_add_par("Decreasing away from stream inlet", style = "heading 2") %>%
  body_add_flextable(soil.decrease.tab) %>%
  body_end_section_landscape() %>%
  print(target = "tables/soil-core-change-tables.docx")
```

```
## [1] "/Users/nawis/GitHub/ReservoirGradient/tables/soil-core-change-tables.docx"
```

Word Table

```
soil.tab <- core.soil %>% arrange(desc(RNA.max)) %>% flextable() %>% autofit()
water.tab <- core.water %>% arrange(desc(RNA.max)) %>% flextable() %>% autofit()

read_docx() %>%
  body_add_par("Table S1", style = "heading 1") %>%
  body_end_section_continuous() %>%
  body_add_par("Core Reservoir Microbiome (present in soils)", style = "heading 2") %>%
  body_add_flextable(soil.tab) %>%
  body_add_par("Core Reservoir Microbiome (absent from soils)", style = "heading 2") %>%
  body_add_flextable(water.tab) %>%
  body_end_section_landscape() %>%
  print(target = "tables/core_tables.docx")
```

```
## [1] "/Users/nawis/GitHub/ReservoirGradient/tables/core_tables.docx"
```

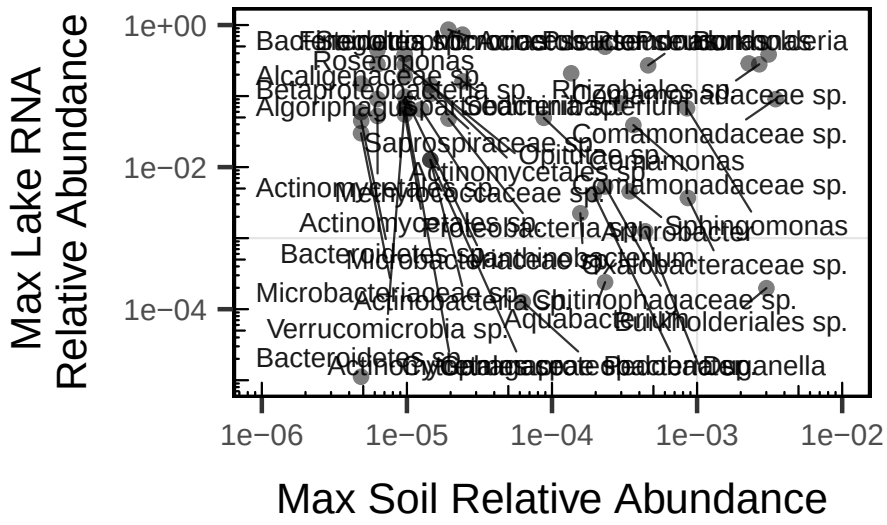
Figure 5: Soil vs. Lake Comparisons

```
soil.vs.lake.abunds %>%
  mutate(Genus = str_replace(Genus, "_unclassified", " sp.")) %>%
  filter(max_soil_relabund > 0) %>%
  ggplot(aes(x = max_soil_relabund, y = RNA.max)) +
  geom_vline(xintercept = 1e-3, alpha = 0.1) +
```

```
  geom_hline(yintercept = 1e-3, alpha = 0.1) +
  geom_jitter(size = 3, alpha = 0.5, show.legend = F) +
  scale_x_log10(lim = c(1e-6, 1e-2)) +
  scale_y_log10(lim = c(1e-5, 1)) +
  annotation_logticks(long = unit(.1, "in")) +
  scale_color_manual(values = my.cols) +
  labs(x = "Max Soil Relative Abundance", y = "Max Lake RNA \nRelative Abundance") +
  geom_text_repel(size = 4.5, aes(label = Genus), force = 1.5, alpha = 0.9, segment.alpha = 0.8, box.pa
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 rows containing missing values (geom_text_repel).
```



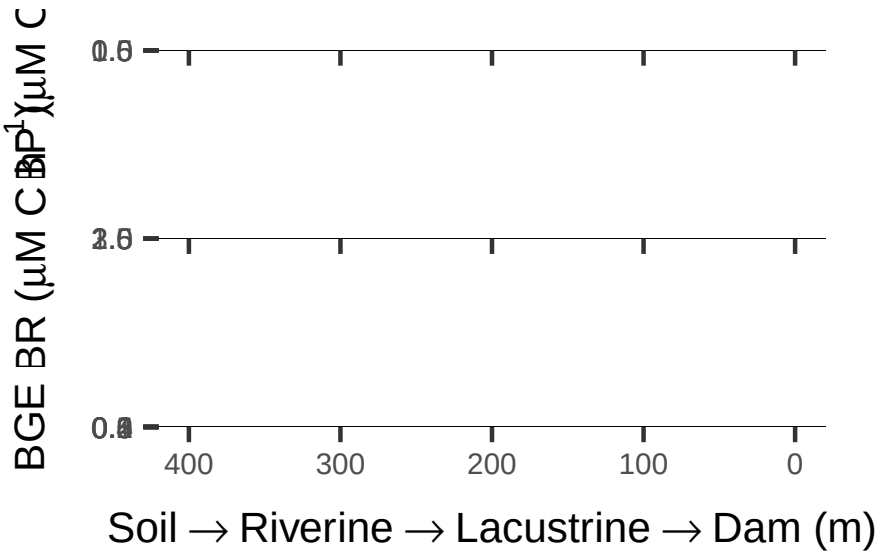
Ecosystem Functioning

Fig 1: Microbial metabolism along reservoir gradient

Read in data

```
metab <- read.table("data/res.grad.metab.txt", sep="\t", header=TRUE)
colnames(metab) <- c("dist", "BP", "BR")
BGE <- round((metab$BP/(metab$BP + metab$BR)),3)
metab <- cbind(metab, BGE)


# Quadratic regression for BP
dist <- metab$dist
dist2 <- metab$dist^2
BP.fit <- lm(metab$BP ~ dist + dist2)
BP.R2 <- round(summary(BP.fit)$r.squared, 2)

# Simple linear regression for BR
BR.fit <- lm(metab$BR ~ metab$dist)
BR.R2 <- round(summary(BR.fit)$r.squared, 2)
BR.int <- BR.fit$coefficients[1]
BR.slp <- BR.fit$coefficients[2]
```

```r
# Simple linear regression for BGE
BGE.fit <- lm(metab$BGE ~ metab$dist)
BGE.R2 <- round(summary(BGE.fit)$r.squared, 2)
BGE.int <- BGE.fit$coefficients[1]
BGE.slp <- BGE.fit$coefficients[2]

BP.R2
BR.R2
BGE.R2

BP.plot <- ggplot(metab, aes(x = dist, y = BP)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x + I(x^2), color = "black") +
  annotate(geom = "text", x = 50, y = 1.5, size = 5, label = paste0("R^2== ",BP.R2), parse = T) +
  labs(y = expression(paste('BP (', mu ,'M C h'^-1* ')')),
       x = (expression("Soil" %->% "Riverine" %->% "Lacustrine" %->% "Dam (m)"))) +
  scale_x_reverse(limits = c(400,0))
BR.plot <- ggplot(metab, aes(x = dist, y = BR)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, color = "black") +
  annotate("text", x = 50, y = 1.5, size = 5, label = paste0("R^2== ",BR.R2), parse = T ) +
  labs(y = expression(paste('BR (', mu ,'M C h'^-1* ')')),
       x = (expression("Soil" %->% "Riverine" %->% "Lacustrine" %->% "Dam (m)"))) +
  scale_x_reverse(limits = c(400,0))
BGE.plot <- ggplot(metab, aes(x = dist, y = BGE)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x + I(x^2), color = "black") +
  annotate("text", x = 50, y = .5, size = 5, label = paste0("R^2== ",BGE.R2), parse = T ) +
  labs(y = "BGE",
       x = (expression("Soil" %->% "Riverine" %->% "Lacustrine" %->% "Dam (m)"))) +
  scale_x_reverse(limits = c(400,0))
plot_grid(BP.plot + theme(axis.title.x = element_blank(), axis.text.x = element_blank(),
                          plot.margin = unit(c(1, 1, -1, 0), "cm")),
          BR.plot + theme(axis.title.x = element_blank(), axis.text.x = element_blank(),
                          plot.margin = unit(c(-1, 1, -1, 0), "cm")),
          BGE.plot + theme(plot.margin = unit(c(-1, 1, 0, 0), "cm")), align = "hv", ncol = 1) +
  ggsave("figures/06_ecosystem-functions.pdf", bg = "white", width = 6, height = 11)
```



Relation of ecosystem functions and community structure

```
# detrend the spatial signal
bp.resid <- resid(lm(BP ~ dist + I(dist)^2, data = metab))
br.resid <- resid(lm(BR ~ dist, data = metab))


metab.resids <- metab
metab.resids$BR_resid <- br.resid + mean(metab$BR)
metab.resids$BP_resid <- bp.resid + mean(metab$BP)

transient.metabolism <- data.frame(transients = terr.REL, dist = design.dna$distance) %>%
  left_join(metab.resids)


bp.mod.quad <- lm(BP_resid ~ transients + I(transients^2), data = transient.metabolism)
bp.mod.lin <- lm(BP_resid ~ transients, data = transient.metabolism)
bp.mod.int <- lm(BP_resid ~ 1, data = transient.metabolism)
anova(bp.mod.int, bp.mod.lin, bp.mod.quad)
AIC(bp.mod.quad, bp.mod.lin, bp.mod.int)

br.mod.quad <- lm(BR_resid ~ transients + I(transients^2), data = transient.metabolism)
br.mod.lin <- lm(BR_resid ~ transients, data = transient.metabolism)
br.mod.int <- lm(BR_resid ~ 1, data = transient.metabolism)
anova(br.mod.int, br.mod.lin, br.mod.quad)
AIC(br.mod.int, br.mod.lin, br.mod.quad)

bge.mod.quad <- lm(BGE ~ transients + I(transients^2), data = transient.metabolism)
bge.mod.lin <- lm(BGE ~ transients, data = transient.metabolism)
bge.mod.int <- lm(BGE ~ 1, data = transient.metabolism)
anova(bge.mod.int, bge.mod.lin, bge.mod.quad)
AIC(bge.mod.int, bge.mod.lin, bge.mod.quad)

round(summary(br.mod.quad)$r.squared, 2)
```

```
round(summary(bp.mod.quad)$r.squared, 2)


total_core <- rowSums(OTUsREL[design$molecule == "DNA" & design$type == "water",
                              subset(rbind.data.frame(high.activity.water.core,
                                                      high.activity.soil.core), RNA.max > .01)$OTU])

summary(lm(BP ~ transients * dist, transient.metabolism))
summary(lm(BR ~ transients * dist, transient.metabolism))


data.frame(
  soil_core = rowSums(OTUsREL[design$molecule == "DNA" & design$type == "water",
           subset(soil.vs.lake.abunds, RNA.max > .01)$OTU]),
  dist = design.dna$distance) %>%
  left_join(metab.resids) %>% select(-BGE, -BP, -BR) %>% gather(metab, value, -soil_core, -dist) %>%
  ggplot(aes(x = soil_core, y = value, color = metab, fill = metab)) +
  geom_point(size = 2) +
  geom_smooth(alpha = .25, method = 'lm', formula = y ~ x + I(x^2)) +
  labs(x = "Relative Abundance of Soil-derived Core",
       y = expression(paste('Metabolism (', mu ,'M C h'^-1* ')'))) +
  scale_color_viridis("Ecosystem Function", discrete = T, begin = .1, end = .6, option = "D") +
  scale_fill_viridis("Ecosystem Function", discrete = T, begin = .1, end = .6, option = "D") +
  ggsave("figures/06_soilcore-function.pdf", bg = "white", width = 7, height = 6)

data.frame(
  water_core = rowSums(OTUsREL[design$molecule == "DNA" & design$type == "water",
                               subset(high.activity.water.core, RNA.max > .01)$OTU]),
  dist = design.dna$distance) %>%
  left_join(metab.resids) %>% select(-BGE,-BR,-BP) %>% gather(metab, value, -water_core, -dist) %>%
  ggplot(aes(x = water_core, y = value, color = metab, fill = metab)) +
  geom_point(size = 2) +
  geom_smooth(alpha = .25, method = 'lm', formula = y ~ x + I(x^2)) +
  labs(x = "Relative Abundance of non-soil-derived Core",
       y = expression(paste('Metabolism (', mu ,'M C h'^-1* ')'))) +
  scale_color_viridis("Ecosystem Function", discrete = T, begin = .1, end = .6, option = "D") +
  scale_fill_viridis("Ecosystem Function", discrete = T, begin = .1, end = .6, option = "D") +
  ggsave("figures/06_nonsoilcore-function.pdf", bg = "white", width = 7, height = 6)



data.frame(transients = resid(lm(terr.REL ~ design.dna$distance)) + mean(terr.REL), dist = design.dna$di
  left_join(metab.resids) %>% select(-BGE, -BP, -BR) %>% gather(metab, value, -transients, -dist) %>%
  ggplot(aes(x = transients, y = value, color = metab, fill = metab)) +
  geom_point(size = 2, show.legend = F) +
  geom_smooth(alpha = .25, method = 'lm', formula = y ~ x, show.legend = F) +
  annotation_logticks(sides = "b") +
  labs(x = "Relative Abundance of Transient Taxa",
       y = expression(paste('Metabolism (', mu ,'M C h'^-1* ')'))) +
  scale_color_viridis("Ecosystem Function", discrete = T, begin = .1, end = .6, option = "D") +
  scale_fill_viridis("Ecosystem Function", discrete = T, begin = .1, end = .6, option = "D") +
  scale_y_continuous(limits = c(0,3)) +
  theme(plot.margin = unit(c(1,1,0,0), "cm")) +
```

```
  ggsave("figures/06_transients-function.pdf", bg = "white", width = 7, height = 6)


core.metab <- data.frame(
  total_core = rowSums(OTUsREL[design$molecule == "DNA" & design$type == "water",
                               subset(rbind.data.frame(high.activity.water.core,
                                                       high.activity.soil.core), RNA.max > .01)$OTU]),
  dist = design.dna$distance) %>%
  left_join(metab.resids)

summary(lm(BP ~ total_core * dist, core.metab))
summary(lm(BR ~ total_core + dist, core.metab))


core.metab <- data.frame(
  total_core = rowSums(OTUsREL[design$molecule == "DNA" & design$type == "water",
                               subset(rbind.data.frame(high.activity.water.core,
                                                       high.activity.soil.core), RNA.max > .01)$OTU]),
  dist = design.dna$distance) %>%
  left_join(metab.resids)
core.metab$total_core_resid <- resid(lm(total_core ~ dist + I(dist^2), core.metab)) + mean(core.metab$t
summary(lm(BP_resid ~ total_core, core.metab))
summary(lm(BR_resid ~ total_core + I(total_core^2), core.metab))


core.metab %>% select(-BGE, -BP, -BR, -total_core) %>% gather(metab, value, -total_core_resid, -dist) %
  ggplot(aes(x = total_core_resid, y = value, color = metab, fill = metab)) +
  geom_point(size = 2, show.legend = F) +
  geom_smooth(alpha = .25, method = 'lm', formula = y ~ x, show.legend = F) +
  labs(x = "Relative Abundance of Core Taxa",
       y = expression(paste('Metabolism (', mu ,'M C h'^-1* ')'))) +
  scale_color_viridis("Ecosystem Function", discrete = T, begin = .1, end = .6, option = "D") +
  scale_fill_viridis("Ecosystem Function", discrete = T, begin = .1, end = .6, option = "D") +
  scale_y_continuous(limits = c(0,3)) +
  theme(plot.margin = unit(c(1,1,0,0), "cm")) +
  ggsave("figures/06_core-function.pdf", bg = "white", width = 7, height = 6)
```

Analyze environmental controls

```
# library(AEM)
# # pull out the in-lake samples
# in.lake.dna.samples <- which(design$type == "water" & design$molecule == "DNA" & design$distance < 30
# env <- env.dat[which(env.dat$sample.ID %in% design$station[in.lake.dna.samples]),c("temp", "pH", "DO"
# env <- scale(env)
# geo_dist <- as.vector(dist(design[in.lake.dna.samples,]$distance))
# env_dist <- as.vector(dist(env))
# com_dist <- as.vector(vegdist(decostand(lake.tot, "total")))
# data.frame(env_dist = resid(lm(env_dist ~ geo_dist)),
#            com_dist = com_dist,
#            geo_dist = resid(lm(geo_dist ~ env_dist))) %>%
#   ggplot(aes(x = env_dist, y = com_dist)) +
```

```
#   geom_point(alpha = 0.5) +
#   geom_smooth(method = 'lm')
# data.frame(env_dist = resid(lm(env_dist ~ geo_dist)),
#            com_dist = com_dist,
#            geo_dist = resid(lm(geo_dist ~ env_dist))) %>%
#   ggplot(aes(x = geo_dist, y = com_dist)) +
#   geom_point(alpha = 0.5) +
#   geom_smooth(method = 'lm')
# data.frame(env_dist = resid(lm(env_dist ~ geo_dist)),
#            com_dist = com_dist,
#            geo_dist = resid(lm(geo_dist ~ env_dist))) %>%
#   ggplot(aes(x = env_dist, y = geo_dist)) +
#   geom_point(alpha = 0.5) +
#   geom_smooth(method = 'lm')
#
# # construct asymetric eigenvector maps along transect
# n <- nrow(env)
# aem.out <- aem.time(n, moran = T, plot.moran = T)
# colnames(aem.out$aem) <- paste0("AEM", seq(1, n-1))
# aems <- aem.out$aem[,which(aem.out$Moran$p.value < 0.05)]
#
#
# prcomp(env, scale. = T)
# round(cor(cbind(aems, env)),2)
# lake.tot <- OTUs[in.lake.dna.samples,]
#
# vp.tot.pos <- varpart(lake.tot, aems[,c("AEM1", "AEM2", "AEM3")], env, transfo = "hellinger")
# vp.tot.pos
# plot(vp.tot.pos)
#
# vp.tot.neg <- varpart(lake.tot, aems[,c("AEM8", "AEM9", "AEM10")], env, transfo = "hellinger")
# vp.tot.neg
# plot(vp.tot.neg)
#
# # rna
# in.lake.rna.samples <- which(design$type == "water" & design$molecule == "RNA" & design$distance < 30
# env <- env.dat[which(env.dat$sample.ID %in% design$station[in.lake.rna.samples]),c("temp", "pH", "DO"
# env <- scale(env)
#
# # construct asymetric eigenvector maps along transect
# n <- nrow(env)
# aem.out <- AEM::aem.time(n, moran = T, plot.moran = T)
# colnames(aem.out$aem) <- paste0("AEM", seq(1, n-1))
# aems <- aem.out$aem[,which(aem.out$Moran$p.value < 0.05)]
#
# prcomp(env, scale. = T)
# round(cor(cbind(aems, env)),2)
# lake.act <- OTUs[in.lake.rna.samples,]
#
# geo_dist <- as.vector(dist(design[in.lake.rna.samples,]$distance))
# env_dist <- as.vector(dist(env))
# com_dist <- as.vector(vegdist(decostand(lake.act, "total")))
# data.frame(env_dist = resid(lm(env_dist ~ geo_dist)),
```

```
#            com_dist = com_dist,
#            geo_dist = resid(lm(geo_dist ~ env_dist))) %>%
#   ggplot(aes(x = env_dist, y = com_dist)) +
#   geom_point(alpha = 0.5) +
#   geom_smooth(method = 'lm')
# data.frame(env_dist = resid(lm(env_dist ~ geo_dist)),
#            com_dist = com_dist,
#            geo_dist = resid(lm(geo_dist ~ env_dist))) %>%
#   ggplot(aes(x = geo_dist, y = com_dist)) +
#   geom_point(alpha = 0.5) +
#   geom_smooth(method = 'lm')
# data.frame(env_dist = resid(lm(env_dist ~ geo_dist)),
#            com_dist = com_dist,
#            geo_dist = resid(lm(geo_dist ~ env_dist))) %>%
#   ggplot(aes(x = env_dist, y = geo_dist)) +
#   geom_point(alpha = 0.5) +
#   geom_smooth(method = 'lm')
#
# vp.act.pos <- varpart(lake.act, aems[,c("AEM1", "AEM2", "AEM3")], env, transfo = "hellinger")
# vp.act.pos
# plot(vp.act.pos)
#
# vp.act.neg <- varpart(lake.act, aems[,c("AEM8", "AEM9", "AEM10")], env, transfo = "hellinger")
# vp.act.neg
# plot(vp.act.neg)
```

	Class	Order	DNA		RNA	
			min	max	min	
Otu00001	Comamonadaceae	Comamonadaceae_unclassified	0.00465	0.026	5.1e-05	0.0
Otu00002	Actinomycetales_unclassified	Actinomycetales_unclassified	0.00327	0.127	0	0
Otu00003	Spartobacteria_unclassified	Spartobacteria_unclassified	0.0016	0.06	2.69e-05	0
Otu00005	Chitinophagaceae	Sediminibacterium	0.00155	0.0369	0	0.0
Otu00006	Saprospiraceae	Saprospiraceae_unclassified	0.000158	0.00806	0	0
Otu00008	Actinomycetales_unclassified	Actinomycetales_unclassified	0.000716	0.0288	0	0.0
Otu00009	Pseudomonadaceae	Pseudomonas	0	0.0412	3.1e-05	0
Otu00010	Proteobacteria_unclassified	Proteobacteria_unclassified	0.00297	0.134	4.25e-05	0.0
Otu00011	Betaproteobacteria_unclassified	Betaproteobacteria_unclassified	0.000108	0.0731	5.23e-06	0.0
Otu00012	Comamonadaceae	Comamonadaceae_unclassified	0.00616	0.0186	8.5e-06	
Otu00014	Actinomycetales_unclassified	Actinomycetales_unclassified	0.00108	0.0512	0	0.0
Otu00015	Actinobacteria_unclassified	Actinobacteria_unclassified	0.000363	0.0675	0	0.0
Otu00016	Microbacteriaceae	Microbacteriaceae_unclassified	0.000115	0.0268	0	
Otu00017	Actinomycetales_unclassified	Actinomycetales_unclassified	0.00103	0.0141	0	0
Otu00018	Pseudomonadaceae	Pseudomonas	4.21e-05	0.0328	3.12e-05	0
Otu00019	Cytophagaceae	Cytophagaceae_unclassified	0.000697	0.0844	0	0.0
Otu00020	Alcaligenaceae	Alcaligenaceae_unclassified	0.000777	0.0399	0	
Otu00022	Opitutae_unclassified	Opitutae_unclassified	0.00421	0.0332	5.23e-06	0
Otu00023	Moraxellaceae	Acinetobacter	0	0.00186	1.55e-05	0
Otu00024	Bacteroidetes_unclassified	Bacteroidetes_unclassified	0.000367	0.00679	0	0.0
Otu00025	Microbacteriaceae	Microbacteriaceae_unclassified	0.00233	0.0271	0	0.0
Otu00028	Pseudomonadaceae	Pseudomonas	0	0.0232	5.23e-06	0
Otu00030	Micrococcaceae	Micrococcus	6.84e-05	0.0215	1.56e-05	0
Otu00031	Cyclobacteriaceae	Algoriphagus	0.000735	0.0293	0	0.0
Otu00032	Bacteroidetes_unclassified	Bacteroidetes_unclassified	0.00101	0.0326	0	0
Otu00033	Rhizobiales_unclassified	Rhizobiales_unclassified	0.00136	0.0398	5.16e-06	0
Otu00039	Comamonadaceae	Comamonas	0.000143	0.0142	0	0.0
Otu00040	Acetobacteraceae	Roseomonas	0.00021	0.015	0	
Otu00042	Burkholderiaceae	Burkholderia	0	0.0129	0	0
Otu00045	Oxalobacteraceae	Oxalobacteraceae_unclassified	0.00103	0.0214	0	0.00
Otu00053	Clostridiales_Incertae_Sedis_XI	Finegoldia	0	0.00102	0	0
Otu00057	Methylococcaceae	Methylococcaceae_unclassified	0.000373	0.0179	0	0.0
Otu00059	Micrococcaceae	Arthrobacter	0	0.0435	0	0.00
Otu00063	Verrucomicrobia_unclassified	Verrucomicrobia_unclassified	0.000573	0.0317	0	0.0
Otu00065	Sphingobacteriaceae	Pedobacter	0	0.0344	0	0.0
Otu00069	Xanthomonadaceae	Stenotrophomonas	0	0.000679	0	0
Otu00072	Sphingomonadaceae	Sphingomonas	7.52e-05	0.118	0	0.0
Otu00078	Flavobacteriaceae	Flavobacterium	5.63e-06	0.00306	0	0.00
Otu00081	Flavobacteriaceae	Flavobacterium	0	0.0154	0	0.00
Otu00082	Oxalobacteraceae	Janthinobacterium	0.000957	0.0141	0	0.0
Otu00087	Bradyrhizobiaceae	Bradyrhizobium	7.74e-06	0.000906	0	0.00
Otu00089	Sphingobacteriales_unclassified	Sphingobacteriales_unclassified	3.82e-05	0.0163	0	0.0
Otu00094	Sphingobacteriaceae	Sphingobacterium	0	0.0142	0	0.00
Otu00095	Oxalobacteraceae	Duganella	4.56e-05	0.0269	5.23e-06	0.0
Otu00098	Sphingomonadaceae	Sphingomonadaceae_unclassified	0	0.00101	0	0.000
Otu00118	Comamonadaceae	Comamonadaceae_unclassified	0.00023	0.00495	0	0.0
Otu00144	Methylococcaceae	Methylobacter	0	0.000353	0	8.86
Otu00145	Caulobacteraceae	Phenylobacterium	0	0.00107	0	1.12
Otu00158	Sphingomonadaceae	Sphingomonas	0	0.000484	0	0.000
Otu00162	Aeromonadaceae	Aeromonas	0	0.000611	0	7.07
Otu00279	Rhizobiaceae	Rhizobiaceae_unclassified	7.74e-06	0.00201	0	0.0
Otu00838	Chitinophagaceae	Chitinophagaceae_unclassified	0	0.000162	0	0.000
Otu01248	Subdivision3_unclassified	Subdivision3_unclassified	0	2.87e-05	0	1.78

	Class	Order	DNA		RNA	
			min	max	min	max
Otu00004	Actinomycetales_unclassified	Actinomycetales_unclassified	0.00348	0.0602	0	0.0227
Otu00007	Burkholderiaceae	Polynucleobacter	0.000697	0.0207	0	0.0865
Otu00038	Actinomycetales_unclassified	Actinomycetales_unclassified	0.00153	0.0222	0	0.0986
Otu00080	Bacteroidetes_unclassified	Bacteroidetes_unclassified	1.91e-05	0.0189	0	0.0188
Otu00090	Opitutae_unclassified	Opitutae_unclassified	0	0.00123	0	0.187
Otu00136	Methylococcaceae	Methylomonas	0	0.00192	0	0.0121
Otu00140	Cryomorphaceae	Fluviicola	0	0.000679	0	0.15
Otu00142	Bacteroidetes_unclassified	Bacteroidetes_unclassified	0.000101	0.0053	0	0.0537
Otu00172	Bacteroidetes_unclassified	Bacteroidetes_unclassified	9.55e-06	0.00224	0	0.00231
Otu00173	Bacteria_unclassified	Bacteria_unclassified	0	0.000459	0	0
Otu00532	Bacteroidetes_unclassified	Bacteroidetes_unclassified	0	0.000772	0	0.000571
Otu00633	Nitrosomonadaceae	Nitrosomonas	0	0.000561	0	2.69e-05
Otu01046	Clostridiales_Incertae_Sedis_XI	Anaerococcus	0	6.54e-05	0	1.41e-05
Otu01198	Burkholderiales_unclassified	Burkholderiales_unclassified	0	0.000274	0	4.24e-05