Dormancy and dispersal structure bacterial communities across ecosystem boundaries

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

20 February, 2019

Initial Setup

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

```r
# Import Required Packages
library("png")
library("grid")
library("tidyverse")
library("vegan")
library("xtable")
library("viridis")
library("cowplot")
library("adespatial")
```

```
## Warning: replacing previous import 'RNeXML::slot<-' by 'methods::slot<-'
## when loading 'phylobase'
```

```
## Warning: replacing previous import 'RNeXML::slot' by 'methods::slot' when
## loading 'phylobase'
```

```r
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.

```r
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 = 16),
        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)),
```

```r
        #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 = 14),
        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.

```r
# 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.

```r
# 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.

```r
facet.labs <- c(`chla` = "Chlorophyll-a",
                `color` = "Color",
                `DO` = "Dissolved Oxygen",
                `pH` = "pH",
                `TP` = "Total Phosphorus")


env.dat %>% select(dist.dam, DO, pH, TP, 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", switch = "y",
             labeller = as_labeller(facet.labs)) +
  theme(strip.background = element_blank(),
        strip.text = element_text(size = 14),
        strip.placement = "outside") +
  labs(x = "Reservoir Transect (m)",
       y = "") +
  scale_x_reverse() +
  scale_y_continuous()
```
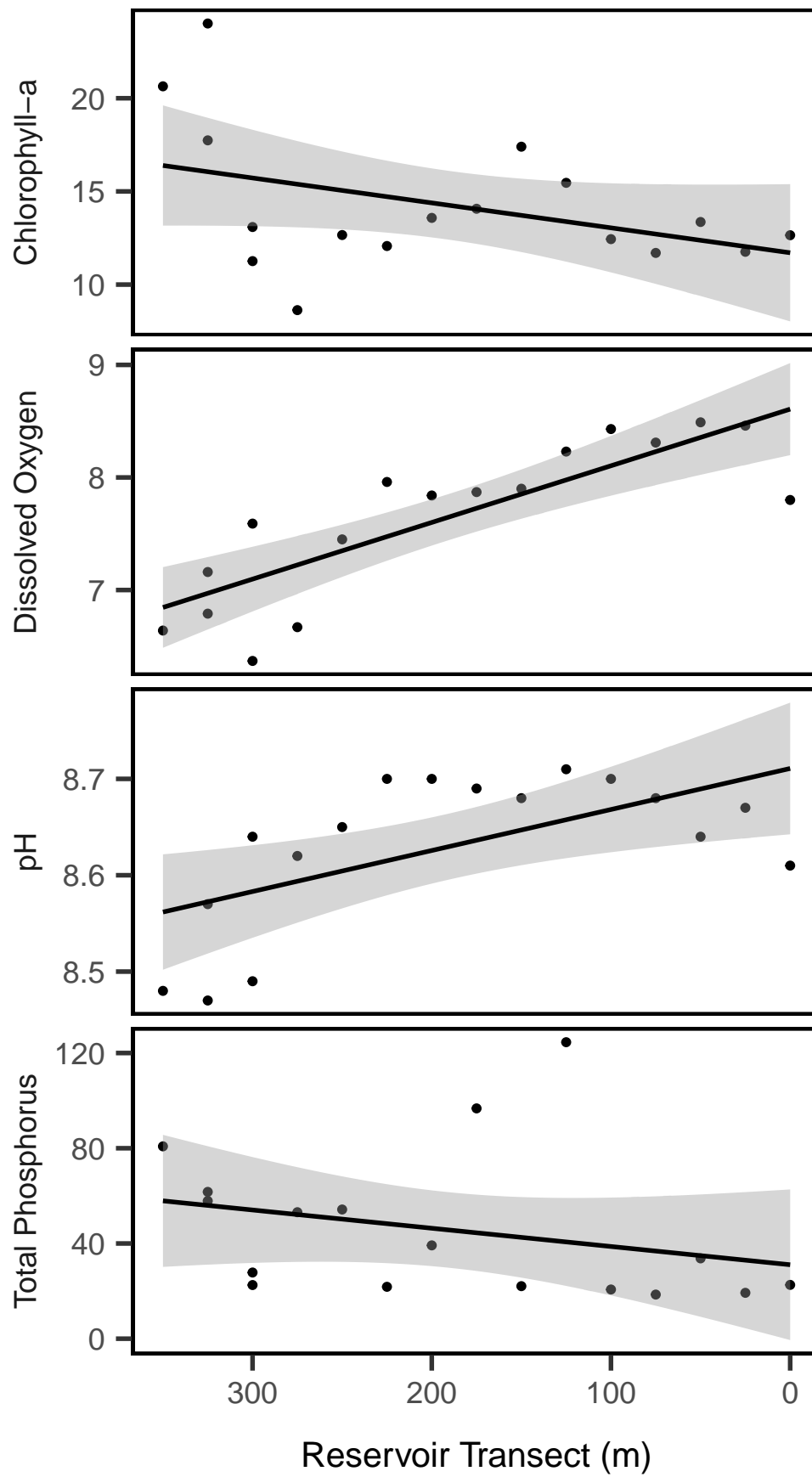
So, there are some weak gradients, but nothing too prevailing.

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.

```r
# 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"]))
```

Here is the resulting curve, showing the higher diversity in soil samples relative to the lake samples.

```r
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 = "Estimated Richness") +
  theme(legend.position = "none") +
  #theme(legend.position =  c(.88,.5)) +
  annotate(label = "Soil", size = 6, geom = "text", x = 450000, y = 7500) +
  annotate(label = "Water", size = 6, geom = "text", x = 450000, y = 1500) +
  scale_color_grey(end = .7) +
  scale_fill_grey(end = .7)
```
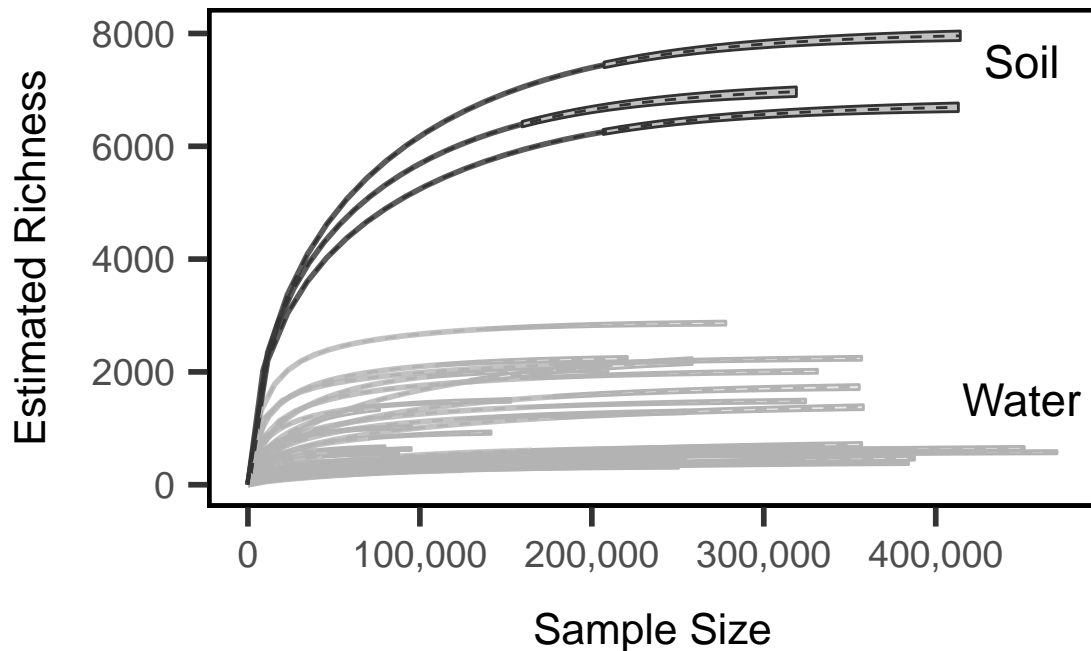
Next, we'll extract the estimates for the Hill numbers at different levels of q, which differentially weight common versus rare species.

```r
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)
# summary(hill.water.mod.shan)
# summary(hill.water.mod.simp)

# tidy up the model output
hill.water.mods <- as_tibble(rbind.data.frame(
  tidy(hill.water.mod.rich) %>% add_column(Diversity = "Richness"),
  tidy(hill.water.mod.shan) %>% add_column(Diversity = "Shannon"),
  tidy(hill.water.mod.simp) %>% add_column(Diversity = "Simpson")
))

# 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()) %>%
  pander(round = 2)
```

Diversity	Term	Estimate	Std. Error	Statistic	p-value
Richness	distance	5.21	0.54	9.66	0
Richness	moleculeRNA	104.9	166.2	0.63	0.53
Richness	distance:moleculeRNA	-5.12	0.72	-7.15	0
Shannon	distance	0.29	0.11	2.75	0.01
Shannon	moleculeRNA	-35.51	32.34	-1.1	0.28
Shannon	distance:moleculeRNA	-0.29	0.14	-2.08	0.05
Simpson	distance	0.05	0.03	1.66	0.11
Simpson	moleculeRNA	-22.19	9.65	-2.3	0.03
Simpson	distance:moleculeRNA	-0.04	0.04	-1.06	0.3

```
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 (Species Equivalents)") +
  scale_color_manual(values = my.cols) +
  scale_fill_manual(values = my.cols) +
  theme(legend.position = c(.88,.95)) +
  scale_x_reverse() +
  facet_wrap(~ Diversity, scales = "free", ncol = 1)
```
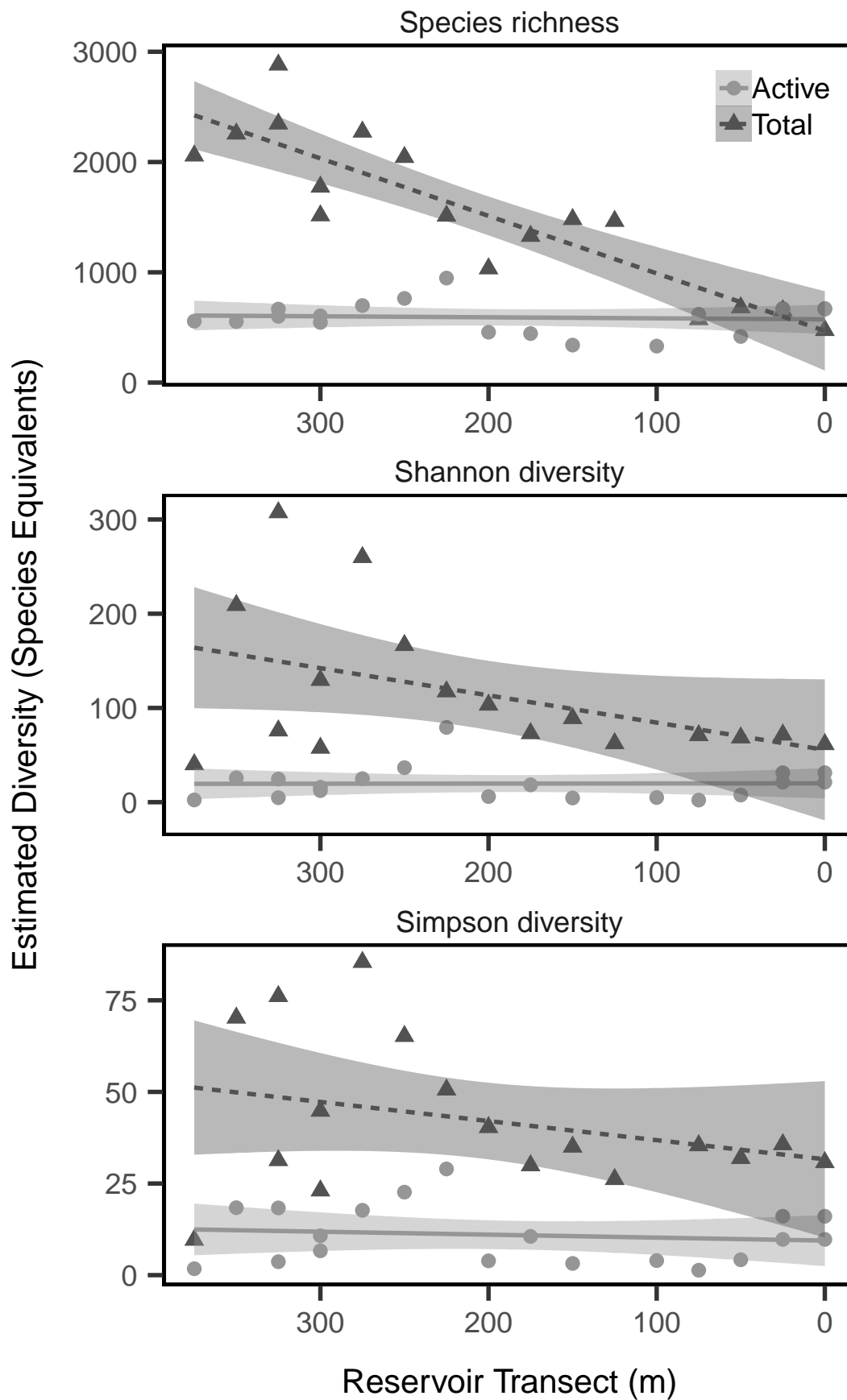
```
    #facet_grid(Diversity ~ ., scales = "free")
```

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")
soil.pcvals <- data.frame(scores(ul.pcoa)) %>%
  rownames_to_column("name") %>%
  left_join(rownames_to_column(design, "name")) %>%
  arrange(desc(distance)) %>% filter(type == "soil")
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_path(size = 1, alpha = 0.75, arrow = arrow(angle = 20,
                          length = unit(0.35, "cm"),
                          type = "closed"), aes(color = molecule, linetype = molecule)) +
  geom_point(size = 3, alpha = 0.8, aes(color = molecule, shape = molecule)) +
  geom_point(data = select(soil.pcvals, Dim1, Dim2), col = "black", alpha = .8, size = 3) +
  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 = -.33, label = "Active", size = 6) +
    annotate(geom = "text", x = -.3, y = .1, label = "Total", size = 6) +
    annotate(geom = "text", x = -.57, y = -.1, label = "Soils", size = 6)
```
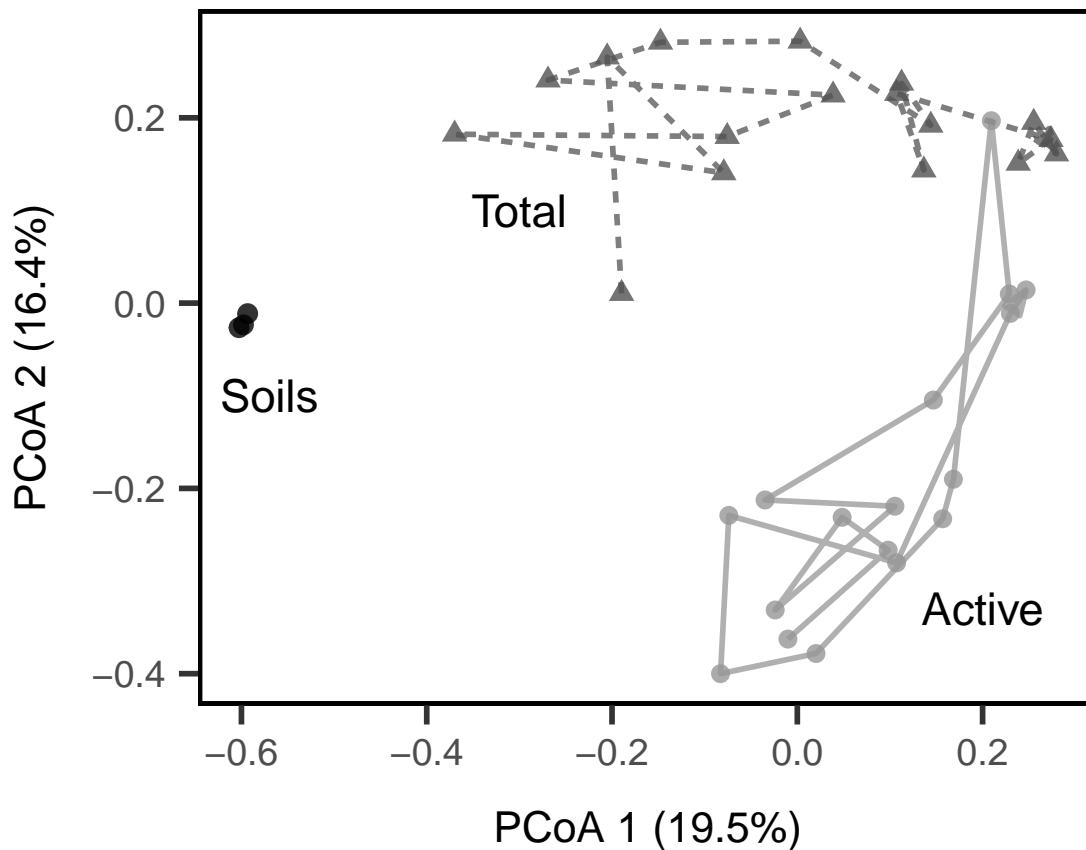

So, it appears that there is convergence in community structure along the path from stream inlet to the dam. This could reflect a loss of soil-derived taxa in the aquatic samples. To test this, we'll look at β-diversity along the gradient with respect to the soil samples. If we see a decay in similarity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

Here, we fit a linear model to the similarity of the aquatic community to the soil community.

```
# 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 2: 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$molecule == "DNA", "Total", "Active")) %>%
  ggplot(aes(x = distance, y = bray.mean,
             color = molecule, fill = molecule, shape = molecule)) +
  geom_point(alpha = 0.8, size = 3, show.legend = T) +
  geom_smooth(method = "lm", show.legend = T, aes(linetype = molecule)) +
  labs(y = str_wrap("Similarity to Soil Community", width = 20),
       x = "Reservoir Transect (m)") +
  scale_color_manual(values = my.cols) +
  scale_fill_manual(values = my.cols) +
  theme(legend.position = c(0.85, 0.85)) +
  scale_x_reverse(limits = c(400,0)) +
  annotate(geom = "text", x = 50, y = 0.15, size = 6,
           label = paste0("R^2== ",round(summary(model.terr)$r.squared, 2)), parse = T)
```
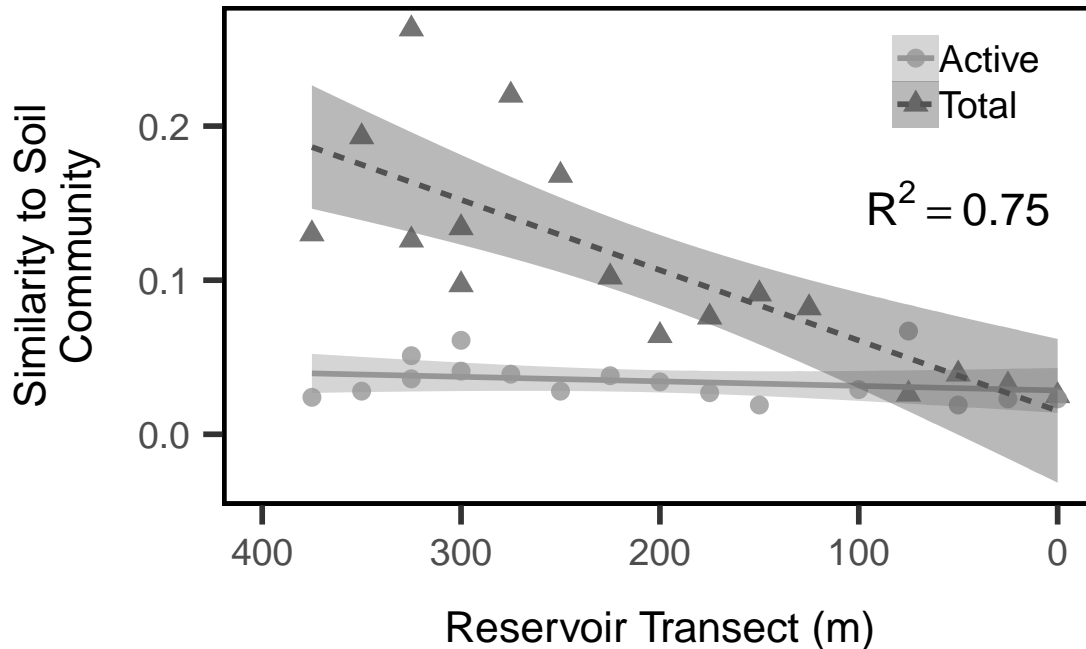
We find that our model captures most of the variation in community structure ($R^2 = 0.74986$). We note a significant influence of distance on community similarity and the presence of a significant interaction between distance and whether the comparison is for active or total bacterial communities. This indicates that total communities decay faster with distance to soils than active communities do, which might be explained by the large difference in initial intercept. Active communities are always highly dissimilar to soil communities and remain so across the lake, while total lake communities are initially similar to soils, but this influence dissipates with distance into the reservoir.

Identifying the Soil Bacteria

Now, we wish to determine whether soil-derived taxa are driving this pattern, and then ask who these influential soil bacteria are.

To classify soil bacteria, we take an incidence-based approach and classify OTUs as:
- present in the soil and present, but never active, in the reservoir
- present in the soil and active in the reservoir

```
# 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 counts for otus found in lake and soil
lake.and.soil.act <- w.rna[,colnames(lake.and.soil.total)]
```

```r
# of these lake and soil taxa, which are never active?
nvr.act <- which(colSums(lake.and.soil.act) == 0)
yes.act <- which(colSums(lake.and.soil.act) != 0)
length(nvr.act) / ncol(lake.and.soil.total)
```

```
## [1] 0.8432942
```

```r
length(yes.act) / ncol(lake.and.soil.total)
```

```
## [1] 0.1567058
```

We calculate the richness of the soil taxa that are never active in the lake. We calculate richness from the DNA-based samples.

```r
# 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$distance)
#summary(terr.mod1)
T1.R2 <- round(summary(terr.mod1)$r.squared, 2)
T1.int <- terr.mod1$coefficients[1]
T1.slp <- terr.mod1$coefficients[2]
pander(terr.mod1)
```
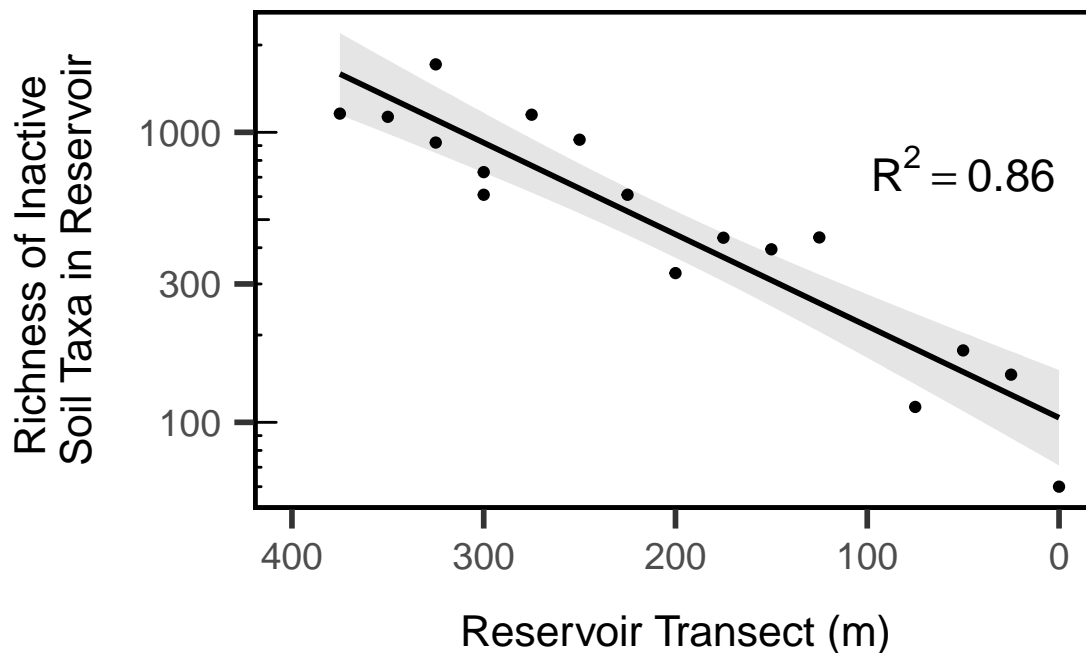
Table 3: Fitting linear model: terr.rich.log ~ design.dna$distance
We find distance is a highly significant predictor of the richness of
these soil-derived taxa (on a log-scale).

	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

```r
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 = "Richness of Inactive \nSoil Taxa in Reservoir") +
  annotate("text", x = 50, y = 750, size = 6, label = paste0("R^2== ",T1.R2), parse = T)
```

The figure shows a scatter plot with axis label "Richness of Inactive Soil Taxa in Reservoir" (y-axis, log scale with values 100, 300, 1000) and "Reservoir Transect (m)" (x-axis, values 400, 300, 200, 100, 0). A negative linear regression line with shaded confidence interval is shown, annotated with $R^2 = 0.86$.

What is the fate of soil-derived taxa in the reservoir?

So, we observe that most soil-derived taxa appear to decay once they enter the reservoir. Do any soil-derived taxa persist in the active bacterial community of the reservoir and do they rise to high relative abundances?

```r
# 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 <- OTUs[which(design$molecule == "RNA" & design$type == "water"), ]
in.lake.rna.pa <- (in.lake.rna > 0) * 1

# define the 'core' taxa as otus present in 50% of samples
in.lake.core <- w.dna[, which((colSums(in.lake.rna.pa) / nrow(in.lake.rna.pa)) >= 0.5)]

# 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))]

# Find the relative abundance of the core taxa and prepare data frame to plot
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, group = 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) +
  scale_y_log10() +
  scale_x_reverse() +
  facet_wrap(~ found, ncol = 1,
             labeller = as_labeller(c(
               `lake` = "Undetected in soils",
               `soils` = "Present in soils")))
```

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

```
## Warning: Removed 301 rows containing non-finite values (stat_smooth).
```
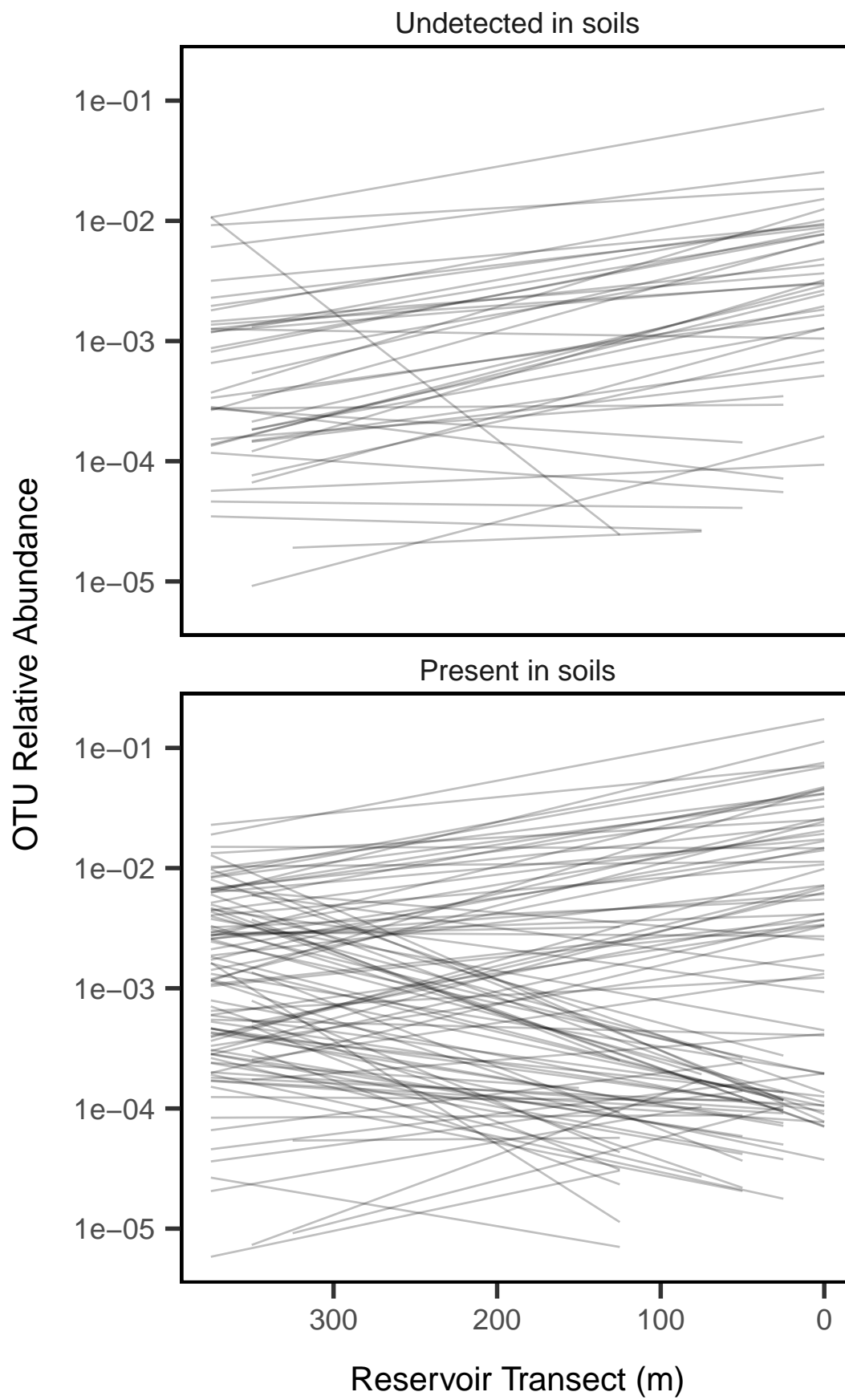
From this figure, we note a few important points. First, we observe that core reservoir taxa that are not detected in the soil samples tend to increase in relative abundance along the reservoir transect. We also note that for the taxa that are present in the soil samples, some tend to increase drastically, while others tend to increase, along the transect. This suggests that there may be two classes of soil-derived OTUs that contribute to reservoir bacterial diversity:

- taxa where the reservoir is a sink (i.e., maintained via mass effects from the soils) - aquatic taxa seeded by populations stored in the soils

```r
# model distance effect on rel abundance to get slope and pval
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")

# classify otus as significantly increasing or decreasing along reservoir
soil.core.decreasing <- 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
```

```r
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
```

```r
nonsoil.core.mods <- apply(in.lake.core.not.soils.REL, MARGIN = 2,
    FUN = function(x) summary(lm(x ~ design.dna$distance))$coefficients[2,c(1,4)])
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
```

```r
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
```

Now we will visualize the significant taxa

```r
pander(nonsoil.core.decreasing, caption = "Core taxa not found in soils that get rarer along the transe
```

Table 4: Core taxa not found in soils that get rarer along the transect. (continued below)

OTU	slope	pval	Domain	Phylum	Class

17

Order	Family	Genus

```
pander(nonsoil.core.increasing, caption = "Core taxa not found in soils that get more common along the
```

Table 6: Core taxa not found in soils that get more common along the transect. (continued below)

OTU	slope	pval	Domain	Phylum
Otu00004	-0.0001388	2.314e-06	Bacteria	Actinobacteria
Otu00038	-4.052e-05	0.0004807	Bacteria	Actinobacteria
Otu00050	-1.946e-05	0.00163	Bacteria	Bacteroidetes
Otu00055	-1.059e-05	0.006773	Bacteria	Bacteroidetes
Otu00058	-1.259e-05	0.01557	Bacteria	Armatimonadetes
Otu00084	-2.095e-06	2.513e-06	Bacteria	Bacteroidetes
Otu00091	-1.471e-05	4.138e-05	Bacteria	Bacteroidetes
Otu00113	-1.147e-06	0.002725	Bacteria	Bacteroidetes
Otu00118	-6.994e-06	0.01551	Bacteria	Actinobacteria
Otu00136	-3.613e-06	0.002073	Bacteria	Verrucomicrobia
Otu00139	-2.359e-06	0.0003476	Bacteria	Bacteroidetes
Otu00150	-1.187e-05	0.008696	Bacteria	Bacteroidetes
Otu00156	-9.037e-06	0.0002416	Bacteria	Bacteria_unclassified
Otu00168	-1.206e-05	0.0007726	Bacteria	Bacteroidetes
Otu00178	-2.593e-06	0.01271	Bacteria	Proteobacteria
Otu00179	-3.867e-06	2.89e-05	Bacteria	Bacteroidetes
Otu00202	-2.737e-06	0.01526	Bacteria	Proteobacteria
Otu00229	-1.215e-05	6.561e-07	Bacteria	Armatimonadetes
Otu00268	-4.396e-06	0.0003199	Bacteria	Planctomycetes
Otu00279	-2.861e-06	0.02887	Bacteria	Proteobacteria
Otu00490	-4.543e-07	0.000452	Bacteria	Proteobacteria
Otu00680	-1.253e-06	0.0004043	Bacteria	Verrucomicrobia

Table 7: Table continues below

Class	Order
Actinobacteria	Actinomycetales
Actinobacteria	Actinomycetales
Sphingobacteriia	Sphingobacteriales
Flavobacteriia	Flavobacteriales
Armatimonadia	Armatimonadales
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Sphingobacteriia	Sphingobacteriales
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Actinobacteria	Actinobacteria_unclassified
Spartobacteria	Spartobacteria_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Bacteria_unclassified	Bacteria_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Alphaproteobacteria	Rhodobacterales
Bacteroidetes_unclassified	Bacteroidetes_unclassified

Class	Order
Betaproteobacteria	Betaproteobacteria_unclassified
Armatimonadetes_gp5	Armatimonadetes_gp5_unclassified
Planctomycetia	Planctomycetales
Gammaproteobacteria	Gammaproteobacteria_unclassified
Betaproteobacteria	Burkholderiales
Spartobacteria	Terrimicrobium

Family	Genus
Actinomycetales_unclassified	Actinomycetales_unclassified
Actinomycetales_unclassified	Actinomycetales_unclassified
Chitinophagaceae	Chitinophagaceae_unclassified
Cryomorphaceae	Cryomorphaceae_unclassified
Armatimonadaceae	Armatimonas/Armatimonadetes_gp1
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Saprospiraceae	Saprospiraceae_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Actinobacteria_unclassified	Actinobacteria_unclassified
Spartobacteria_unclassified	Spartobacteria_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Bacteria_unclassified	Bacteria_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Rhodobacteraceae	Rhodobacteraceae_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Betaproteobacteria_unclassified	Betaproteobacteria_unclassified
Armatimonadetes_gp5_unclassified	Armatimonadetes_gp5_unclassified
Planctomycetaceae	Planctomycetaceae_unclassified
Gammaproteobacteria_unclassified	Gammaproteobacteria_unclassified
Burkholderiaceae	Polynucleobacter
Terrimicrobium_unclassified	Terrimicrobium_unclassified

```
pander(soil.core.decreasing, caption = "Core taxa found in soils that get rarer along the transect.")
```

Table 9: Core taxa found in soils that get rarer along the transect.
(continued below)

OTU	slope	pval	Domain	Phylum
Otu00018	4.776e-05	0.02307	Bacteria	Proteobacteria
Otu00026	1.5e-05	0.03406	Bacteria	Proteobacteria
Otu00057	2.416e-05	0.03413	Bacteria	Proteobacteria
Otu00077	5.224e-05	0.04372	Bacteria	Bacteroidetes
Otu00081	2.098e-05	0.04653	Bacteria	Proteobacteria
Otu00138	3.148e-05	0.04819	Bacteria	Firmicutes
Otu00185	4.128e-06	0.04492	Bacteria	Proteobacteria
Otu00201	1.228e-05	0.04164	Bacteria	Acidobacteria
Otu00242	2.27e-06	0.03891	Bacteria	Proteobacteria
Otu00260	9.499e-06	0.04541	Bacteria	Proteobacteria
Otu00373	6.811e-06	0.03796	Bacteria	Acidobacteria
Otu00545	2.148e-06	0.04557	Bacteria	Actinobacteria

OTU	slope	pval	Domain	Phylum
Otu00718	1.643e-06	0.04315	Bacteria	Proteobacteria
Otu00816	1.938e-06	0.009083	Bacteria	Acidobacteria
Otu00981	1.278e-06	0.01541	Bacteria	Proteobacteria
Otu01104	8.469e-07	0.001233	Bacteria	Acidobacteria
Otu01268	9.938e-07	0.009538	Bacteria	Acidobacteria
Otu01432	5.582e-07	0.009584	Bacteria	Bacteria_unclassified

Table 10: Table continues below

Class	Order
Gammaproteobacteria	Pseudomonadales
Betaproteobacteria	Burkholderiales
Gammaproteobacteria	Methylococcales
Flavobacteriia	Flavobacteriales
Betaproteobacteria	Burkholderiales
Bacilli	Bacillales
Alphaproteobacteria	Rhizobiales
Acidobacteria_Gp6	Gp6
Alphaproteobacteria	Rhizobiales
Gammaproteobacteria	Enterobacteriales
Acidobacteria_Gp6	Gp6
Actinobacteria	Solirubrobacterales
Gammaproteobacteria	Pseudomonadales
Acidobacteria_Gp6	Gp6
Betaproteobacteria	Betaproteobacteria_unclassified
Acidobacteria_Gp6	Gp6
Acidobacteria_Gp6	Gp6
Bacteria_unclassified	Bacteria_unclassified

Family	Genus
Pseudomonadaceae	Pseudomonas
Comamonadaceae	Comamonadaceae_unclassified
Methylococcaceae	Methylococcaceae_unclassified
Flavobacteriaceae	Flavobacterium
Oxalobacteraceae	Janthinobacterium
Bacillaceae_1	Bacillus
Rhizobiales_unclassified	Rhizobiales_unclassified
Gp6_unclassified	Gp6_unclassified
Rhizobiales_unclassified	Rhizobiales_unclassified
Enterobacteriaceae	Yersinia
Gp6_unclassified	Gp6_unclassified
Solirubrobacteraceae	Solirubrobacter
Pseudomonadaceae	Pseudomonas
Gp6_unclassified	Gp6_unclassified
Betaproteobacteria_unclassified	Betaproteobacteria_unclassified
Gp6_unclassified	Gp6_unclassified
Gp6_unclassified	Gp6_unclassified
Bacteria_unclassified	Bacteria_unclassified

```
pander(soil.core.increasing, caption = "Core taxa found in soils that get more common along the transec
```

Table 12: Core taxa found in soils that get more common along the transect. (continued below)

OTU	slope	pval	Domain	Phylum
Otu00001	-2.271e-05	0.02914	Bacteria	Proteobacteria
Otu00002	-0.0002387	0.0004954	Bacteria	Actinobacteria
Otu00003	-0.0001089	0.0004266	Bacteria	Verrucomicrobia
Otu00005	-5.373e-05	0.001508	Bacteria	Bacteroidetes
Otu00006	-9.301e-06	0.02246	Bacteria	Bacteroidetes
Otu00008	-4.31e-05	0.00383	Bacteria	Actinobacteria
Otu00014	-0.0001028	0.000182	Bacteria	Actinobacteria
Otu00015	-0.0001462	4.754e-05	Bacteria	Actinobacteria
Otu00016	-5.778e-05	0.0002333	Bacteria	Actinobacteria
Otu00017	-3.333e-05	2.84e-05	Bacteria	Actinobacteria
Otu00025	-5.211e-05	0.0005492	Bacteria	Actinobacteria
Otu00029	-3.376e-05	0.001502	Bacteria	Actinobacteria
Otu00031	-5.931e-05	0.0002164	Bacteria	Bacteroidetes
Otu00034	-9.673e-06	4.088e-05	Bacteria	Proteobacteria
Otu00040	-3.772e-05	8.166e-06	Bacteria	Proteobacteria
Otu00068	-1.868e-05	0.00012	Bacteria	Bacteroidetes
Otu00071	-5.207e-05	1.123e-05	Bacteria	Planctomycetes
Otu00073	-6.533e-06	0.003648	Bacteria	Proteobacteria
Otu00075	-2.195e-05	0.003707	Bacteria	Bacteria_unclassified
Otu00080	-2.2e-05	0.03337	Bacteria	Bacteroidetes
Otu00096	-7.12e-06	0.004794	Bacteria	Proteobacteria
Otu00099	-2.172e-06	0.02096	Bacteria	Bacteria_unclassified
Otu00101	-9.127e-06	0.0001354	Bacteria	Proteobacteria
Otu00109	-4.411e-06	0.01879	Bacteria	Actinobacteria
Otu00165	-9.645e-06	0.01791	Bacteria	Verrucomicrobia
Otu00190	-3.299e-06	0.01902	Bacteria	Verrucomicrobia
Otu00207	-4.816e-06	0.004282	Bacteria	Proteobacteria

Table 13: Table continues below

Class	Order
Betaproteobacteria	Burkholderiales
Actinobacteria	Actinomycetales
Spartobacteria	Spartobacteria_unclassified
Sphingobacteriia	Sphingobacteriales
Sphingobacteriia	Sphingobacteriales
Actinobacteria	Actinomycetales
Actinobacteria	Actinomycetales
Actinobacteria	Actinobacteria_unclassified
Actinobacteria	Actinomycetales
Actinobacteria	Actinomycetales
Actinobacteria	Actinomycetales
Actinobacteria	Actinomycetales
Cytophagia	Cytophagales
Alphaproteobacteria	Sphingomonadales
Alphaproteobacteria	Rhodospirillales

Class	Order
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Planctomycetia	Planctomycetales
Betaproteobacteria	Burkholderiales
Bacteria_unclassified	Bacteria_unclassified
Flavobacteriia	Flavobacteriales
Alphaproteobacteria	Rhodobacterales
Bacteria_unclassified	Bacteria_unclassified
Alphaproteobacteria	Sphingomonadales
Actinobacteria	Acidimicrobiales
Opitutae	Opitutales
Verrucomicrobiae	Verrucomicrobiales
Alphaproteobacteria	Sphingomonadales

Family	Genus
Comamonadaceae	Comamonadaceae_unclassified
Actinomycetales_unclassified	Actinomycetales_unclassified
Spartobacteria_unclassified	Spartobacteria_unclassified
Chitinophagaceae	Sediminibacterium
Saprospiraceae	Saprospiraceae_unclassified
Actinomycetales_unclassified	Actinomycetales_unclassified
Actinomycetales_unclassified	Actinomycetales_unclassified
Actinobacteria_unclassified	Actinobacteria_unclassified
Microbacteriaceae	Microbacteriaceae_unclassified
Actinomycetales_unclassified	Actinomycetales_unclassified
Microbacteriaceae	Microbacteriaceae_unclassified
Actinomycetales_unclassified	Actinomycetales_unclassified
Cyclobacteriaceae	Algoriphagus
Sphingomonadaceae	Sphingorhabdus
Acetobacteraceae	Roseomonas
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Planctomycetaceae	Planctomycetaceae_unclassified
Comamonadaceae	Polaromonas
Bacteria_unclassified	Bacteria_unclassified
Flavobacteriaceae	Flavobacterium
Rhodobacteraceae	Rhodobacter
Bacteria_unclassified	Bacteria_unclassified
Sphingomonadaceae	Sandarakinorhabdus
Acidimicrobiales_unclassified	Acidimicrobiales_unclassified
Opitutaceae	Opitutus
Verrucomicrobiaceae	Luteolibacter
Erythrobacteraceae	Porphyrobacter

```
# p1 <- 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(Phylum, Class, Order, Family, Genus)) %>%
#   ggplot(aes(x = distance, y = rel_abund, group = OTU)) +
#   #geom_point(alpha = 0.5) +
```

```r
#   geom_line(stat = "smooth", alpha = 0.5, size = 1,
#             color = "black", method = "loess", span = 1, se = FALSE) +
#   scale_x_reverse() +
#   scale_y_log10(labels = scales::scientific) +
#   theme(legend.position = "none") +
#   guides(color = guide_legend(ncol = 1)) +
#   labs(x = "",
#        y = "Relative Abundance",
#        title = "Absent from soil and significantly increasing")
#
# p2 <- as.data.frame(OTUsREL[,soil.core.increasing$OTU]) %>%
#   rownames_to_column("sampleID") %>%
#   left_join(rownames_to_column(design, "sampleID")) %>%
#   gather(OTU, rel_abund, -station, -molecule, -type, -distance, -sampleID) %>%
#   filter(molecule == "DNA") %>% left_join(OTU.tax) %>%
#   mutate(taxon = paste(Class, Order)) %>%
#   ggplot(aes(x = distance, y = rel_abund, group = OTU)) +
#   #geom_point(alpha = 0.5) +
#   geom_line(stat = "smooth", alpha = 0.5, size = 1,
#             color = "black", method = "loess", span = 1, se = FALSE) +
#   scale_x_reverse() +
#   scale_y_log10(labels = scales::scientific) +
#   theme(legend.position = "none") +
#   guides(color = guide_legend(ncol = 1)) +
#   labs(x = "",
#        y = "Relative Abundance",
#        title = "Present in soil and significantly increasing")
#
# p3 <- as.data.frame(OTUsREL[,soil.core.decreasing$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)) %>%
#   ggplot(aes(x = distance, y = rel_abund, group = OTU)) +
#   #geom_point(alpha = 0.5) +
#   geom_line(stat = "smooth", alpha = 0.5, size = 1,
#             color = "black", method = "loess", span = 1, se = FALSE) +
#   scale_x_reverse() +
#   scale_y_log10(labels = scales::scientific) +
#   theme(legend.position = "none") +
#   guides(color = guide_legend(ncol = 1)) +
#   labs(x = "Reservoir Transect (m)",
#        y = "Relative Abundance",
#        title = "Present in soil and significantly decreasing")
#
# cowplot::plot_grid(p1, p2, p3, align = "hv", labels = "AUTO", ncol = 1)

df1 <- 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) %>%
```

```r
  mutate(soils = "Absent from soils", change = "Increasing")
```

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

```r
n1 <- length(unique(df1$OTU))

df2 <- 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(soils = "Present in soils", change = "Increasing")
```

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

```r
n2 <- length(unique(df2$OTU))

df3 <- as.data.frame(OTUsREL[,soil.core.decreasing$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(soils = "Present in soils", change = "Decreasing")
```

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

```r
n3 <- length(unique(df3$OTU))
df.plot <- as_tibble(rbind.data.frame(df1, df2, df3)) %>% filter(type == "water")


df.plot %>% mutate(rel_abund = ifelse(rel_abund == 0, 1e-6, rel_abund)) %>%
  filter(soils == "Present in soils") %>%
  mutate(change = ifelse(change == "Increasing",
                         paste0("Increasing (n = ", n2,")"),
                         paste0("Decreasing (n = ", n3,")"))) %>%
  ggplot(aes(x = distance, y = rel_abund, group = OTU)) +
  #geom_point(alpha = 0.5) +
  geom_line(stat = "smooth", alpha = 0.25, size = 1,
            color = "black", method = "loess", span = 1, se = FALSE) +
  scale_x_reverse() +
  scale_y_log10(labels = scales::scientific) +
  theme(legend.position = "none") +
  guides(color = guide_legend(ncol = 1)) +
  labs(x = "Reservoir Transect (m)",
       y = "Relative Abundance") +
  facet_wrap( ~ change)
```
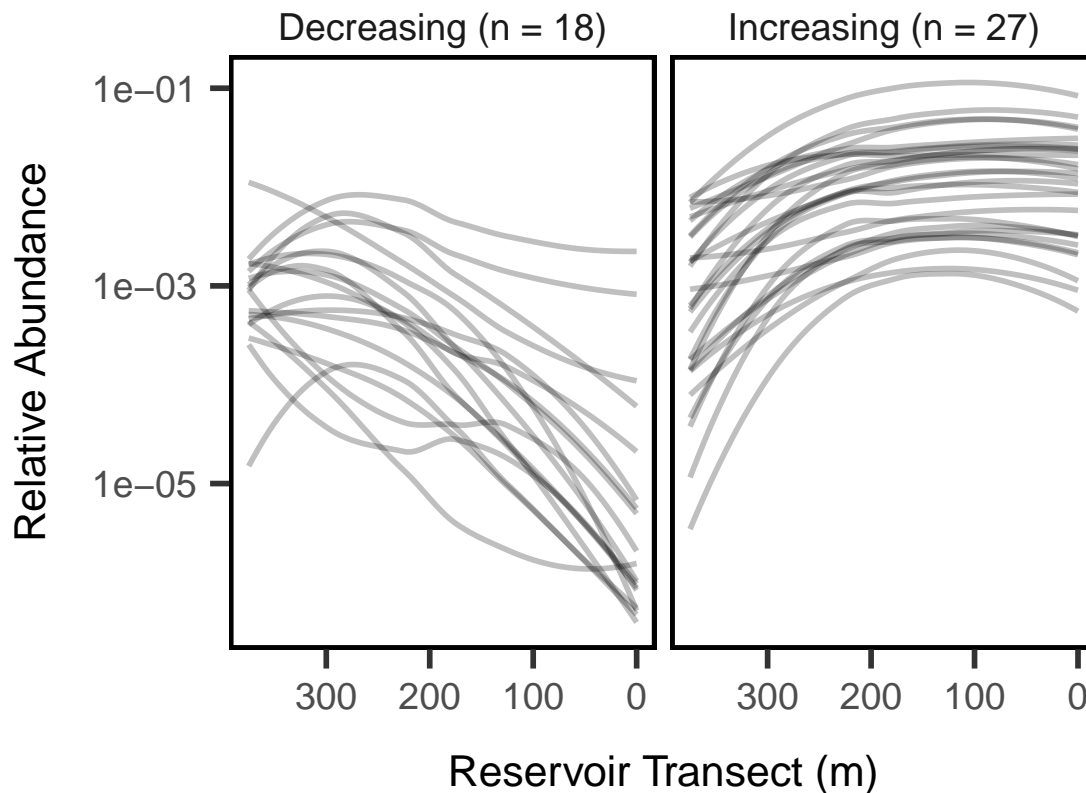
```r
# how much do the different core components contribute to total abundances
in.lake.core.soil.REL <- rowSums(in.lake.core.from.soils) / rowSums(w.dna)
in.lake.core.water.REL <- rowSums(in.lake.core.not.soils) / rowSums(w.dna)
```

Taxonomic Analysis

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

```r
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", "RI




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", "I

# 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)

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.

```r
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)
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)

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)) %>%
```

```r
  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")

# 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))
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))

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")
```

Word Table

```r
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")
```

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
```

Not-included

Ecosystem functions

```
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]

# 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 = "Reservoir Transect (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 = "Reservoir Transect (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 = "Reservoir Transect (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, labels = "AUTO")
```

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)
```

```r
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* ')'))) +
```

```r
  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$total_core_resid <- resid(lm(total_core ~ dist + I(dist^2), core.metab)) + mean(core.metab$to
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")) +
```

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