Dormancy and dispersal structure bacterial communities across ecosystem boundaries

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

28 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)),
        #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.

```
# 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

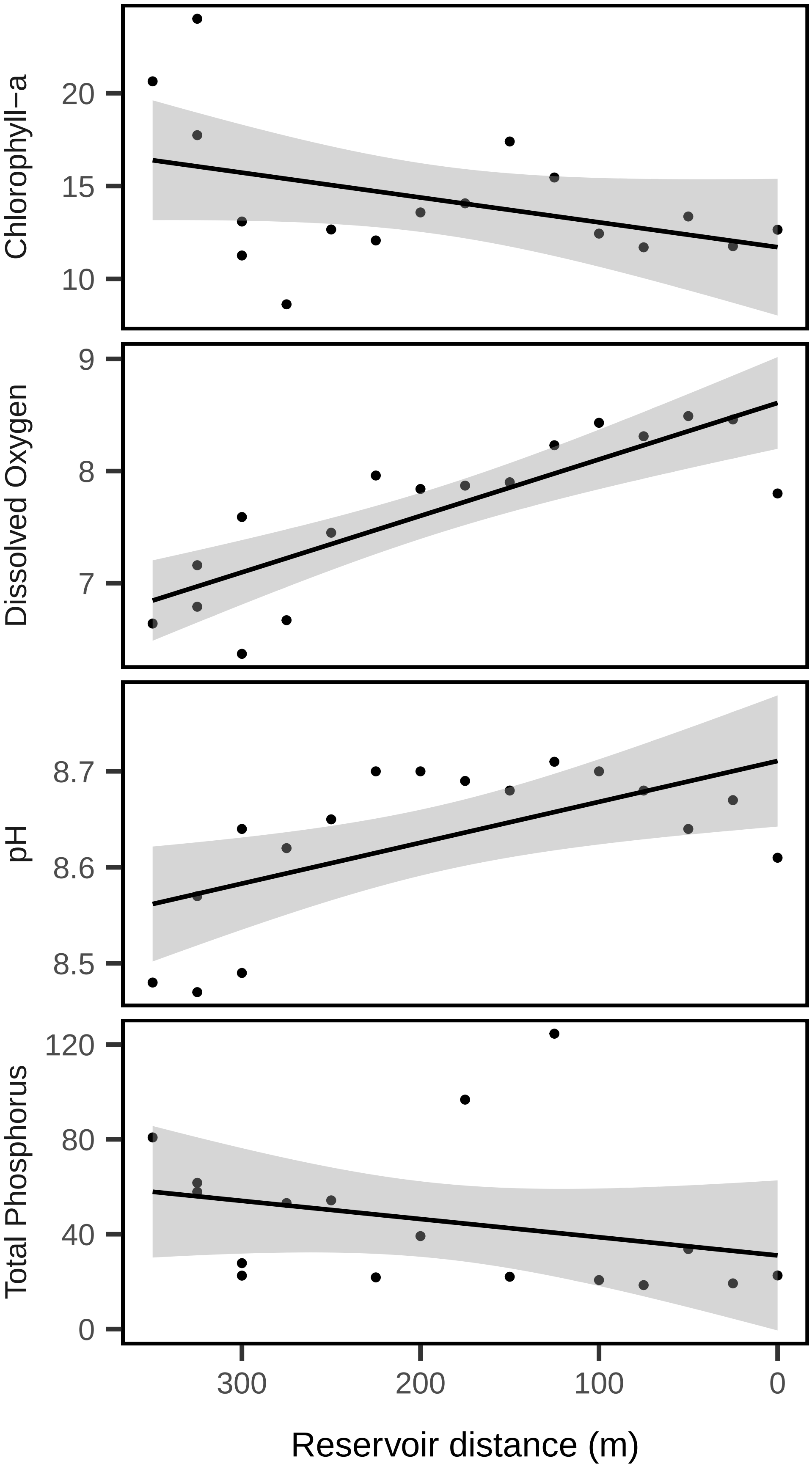
Here, we remove OTUs with low incidence across sites, we remove any samples with low coverage, and we standardize the OTU table by log-transforming the abundances and relativizing by site.

```
# 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)]
coverage <- rowSums(OTUs)
set.seed(47405)
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.

```
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 distance (m)",
       y = "") +
  scale_x_reverse() +
  scale_y_continuous()
```


Chlorophyll−a
10
15
20
Dissolved Oxygen
7
8
9
pH
8.5
8.6
8.7
Total Phosphorus
0
40
80
120
300
200
100
0
Reservoir distance (m)

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.

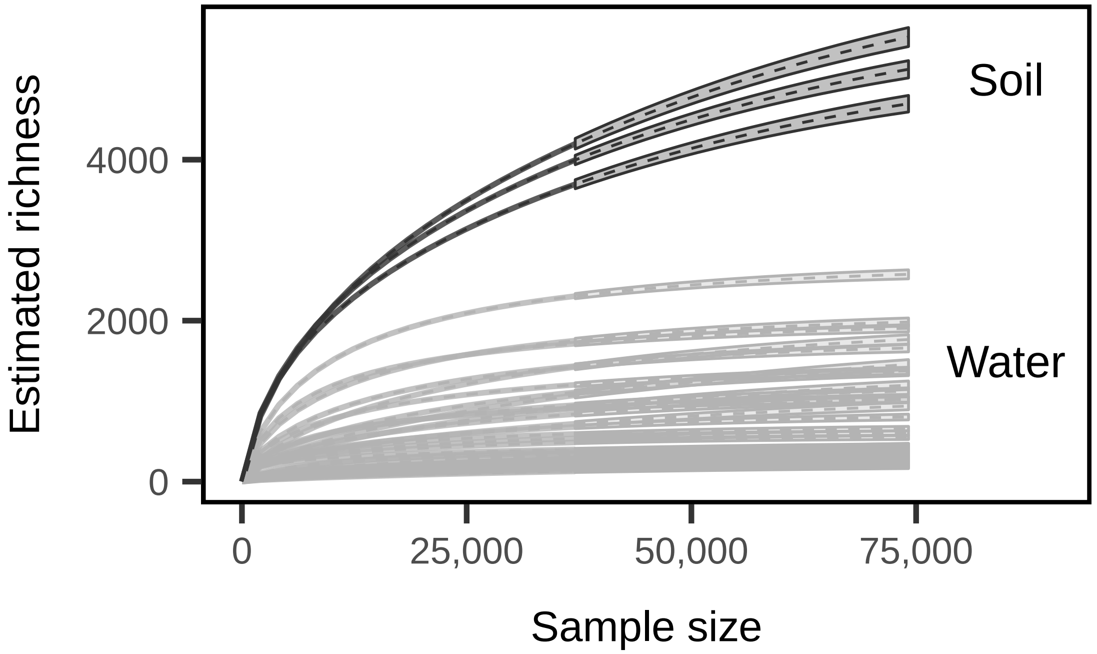
How does α-diversity vary along the reservoir?

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, limits = c(0, 90000)) +
  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 = 85000, y = 5000) +
  annotate(label = "Water", size = 6, geom = "text", x = 85000, y = 1500) +
  scale_color_grey(end = .7) +
  scale_fill_grey(end = .7)
```



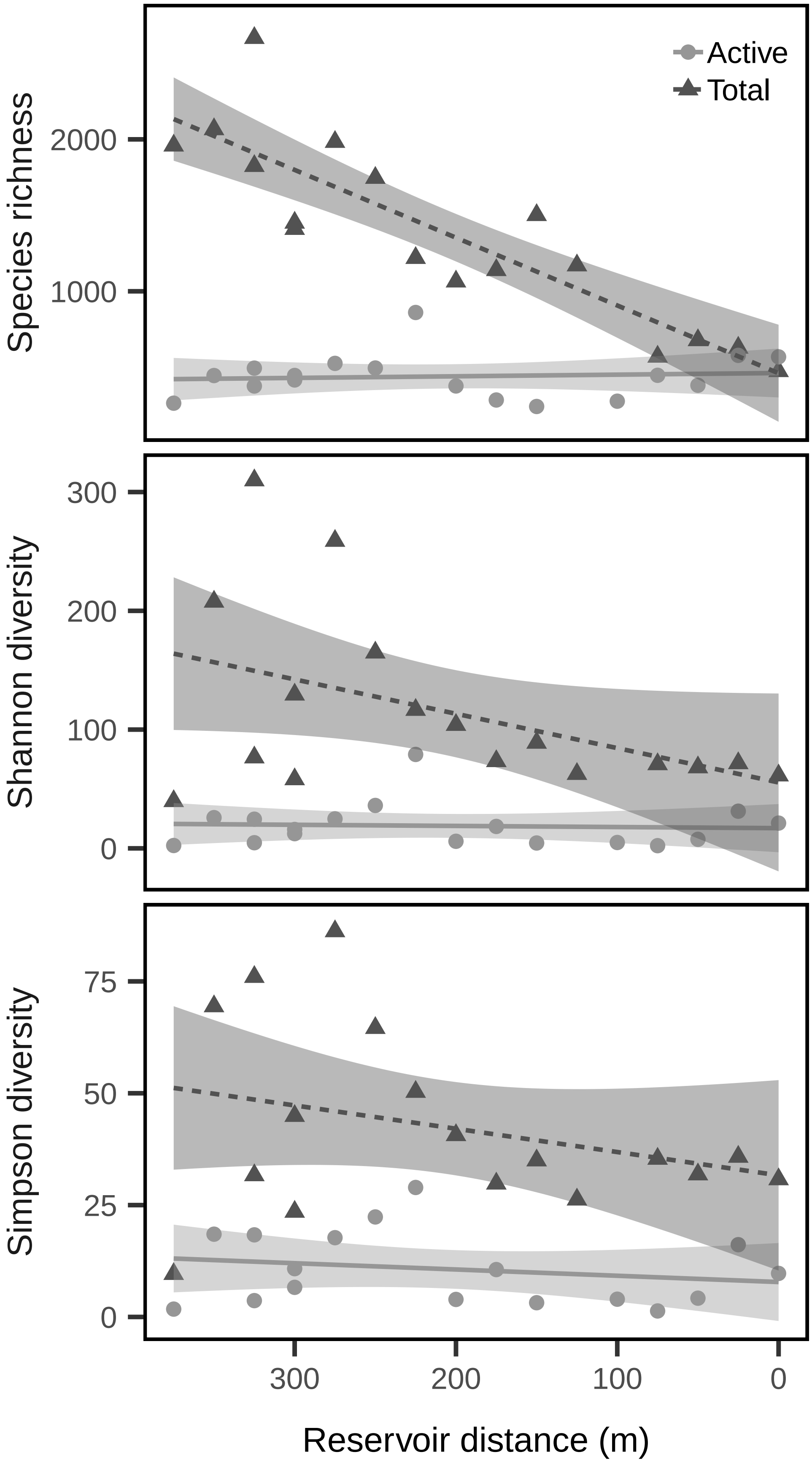
Next, we'll extract the estimates for the Hill numbers at different levels of q, which differentially weight common versus rare species.

```
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 = 4)
```

Diversity	Term	Estimate	Std. Error	Statistic	p-value
Richness	distance	4.461	0.5005	8.912	0
Richness	moleculeRNA	1.364	167.2	0.0082	0.9935
Richness	distance:moleculeRNA	-4.568	0.7043	-6.486	0
Shannon	distance	0.2892	0.1084	2.669	0.0122
Shannon	moleculeRNA	-38.48	36.2	-1.063	0.2963
Shannon	distance:moleculeRNA	-0.2798	0.1525	-1.835	0.0765
Simpson	distance	0.0521	0.0322	1.62	0.1158
Simpson	moleculeRNA	-23.84	10.74	-2.22	0.0341
Simpson	distance:moleculeRNA	-0.0381	0.0453	-0.8415	0.4067

```
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 distance (m)",
       y = "") +
  scale_color_manual(values = my.cols) +
  scale_fill_manual(values = my.cols) +
  theme(legend.position = c(.88,.95), strip.placement = "outside",
        strip.text = element_text(size = 16)) +
  scale_x_reverse() +
  facet_grid(Diversity ~ ., scales = "free", switch = "y") +
  guides(fill = guide_legend(override.aes=list(fill=NA)))
```


Active
Total
Species richness
2000
1000
Shannon diversity
300
200
100
0
Simpson diversity
75
50
25
0
300
200
100
0
Reservoir distance (m)

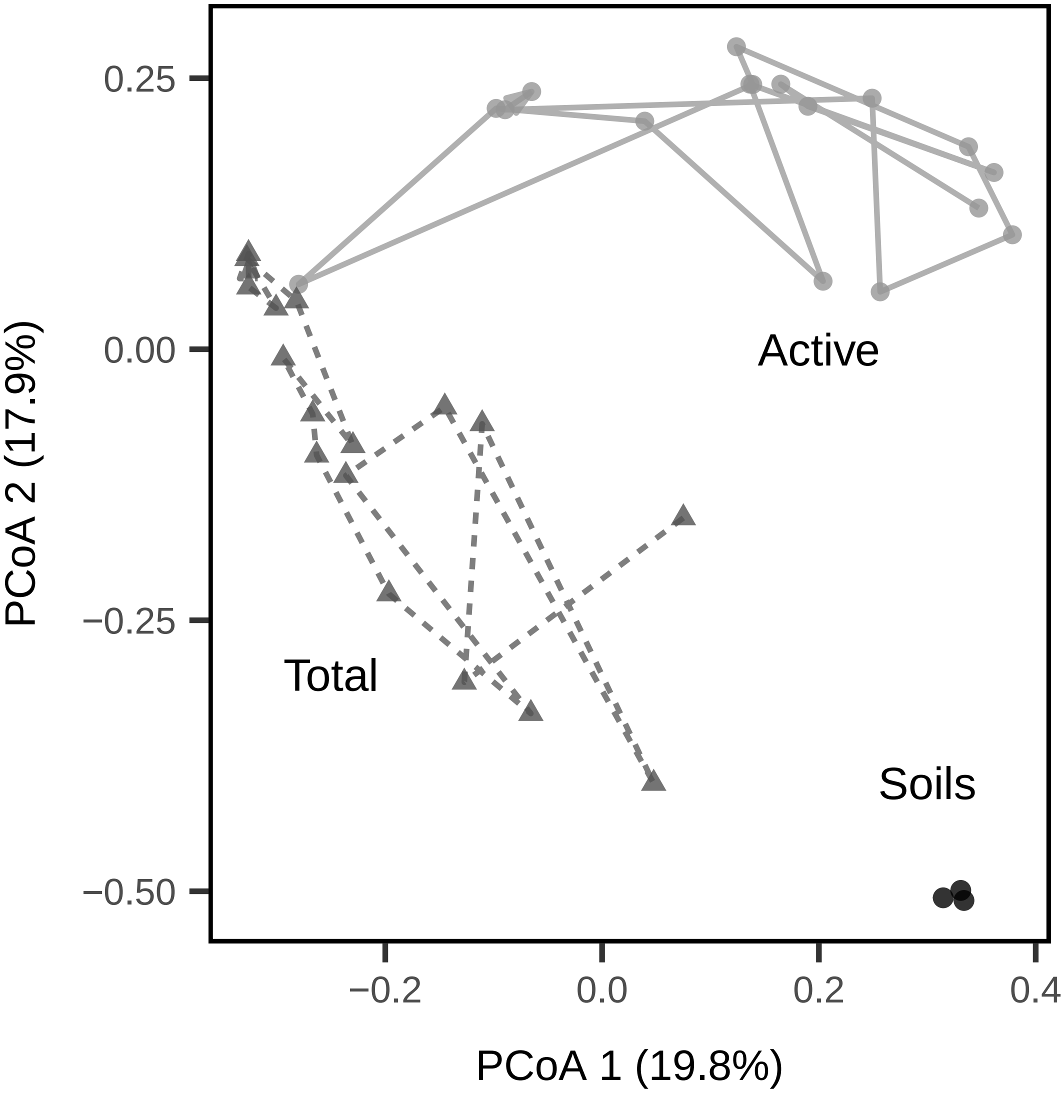
```
  #facet_grid(Diversity ~ ., scales = "free")
```

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 = 0, label = "Active", size = 6) +
  annotate(geom = "text", x = -.25, y = -.3, label = "Total", size = 6) +
  annotate(geom = "text", x = .3, y = -.4, 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)
```

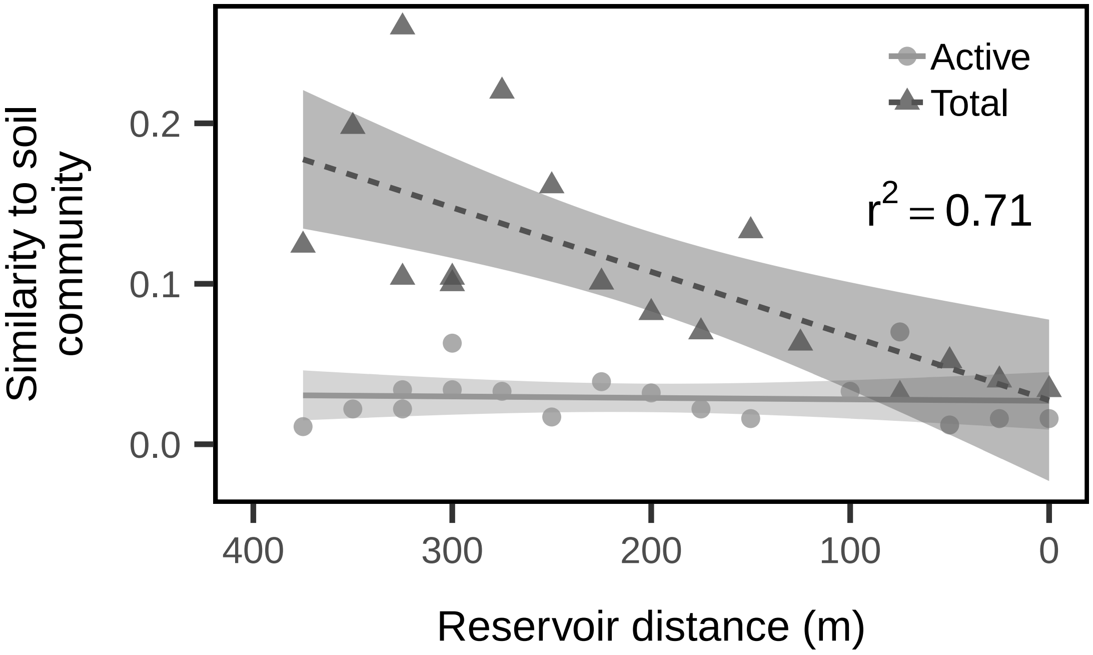
```
# Calculate Linear Model
model.terr <- lm(bray.mean ~ distance * molecule, data = UL.sim)
pander(model.terr)
```

Table 2: Fitting linear model: bray.mean ~ distance * molecule

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.02739	0.01774	1.544	0.1331
distance	0.0004004	7.464e-05	5.365	8.319e-06
moleculeRNA	-0.0003186	0.02493	-0.01278	0.9899
distance:moleculeRNA	-0.0003913	0.000105	-3.726	0.000806

```
# # 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$molecule == "DNA", "Total", "Active")) %>%
  ggplot(aes(x = distance, y = bray.mean,
             color = molecule, fill = molecule, shape = molecule)) +
  geom_point(alpha = 0.8, size = 3, show.legend = T) +
  geom_smooth(method = "lm", show.legend = T, aes(linetype = molecule)) +
  labs(y = str_wrap("Similarity to soil community", width = 20),
       x = "Reservoir distance (m)") +
  scale_color_manual(values = my.cols) +
  scale_fill_manual(values = my.cols) +
  theme(legend.position = c(0.85, 0.85)) +
  scale_x_reverse(limits = c(400,0)) +
  annotate(geom = "text", x = 50, y = 0.15, size = 6,
           label = paste0("r^2== ",round(summary(model.terr)$r.squared, 2)), parse = T) +
  guides(fill = guide_legend(override.aes=list(fill=NA)))
```



We find that our model captures most of the variation in community structure ($R^2 = 0.7084136$). 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

```r
# 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? active?
nvr.act <- which(colSums(lake.and.soil.act) == 0)
yes.act <- which(colSums(lake.and.soil.act) != 0)

# how many otus are active relative to the total number of otus
length(nvr.act) / ncol(lake.and.soil.total)
```

```
## [1] 0.8814706
```

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

```
## [1] 0.1185294
```

```r
# of taxa who were never active, what fraction of the total community did they represent?
sum(rowSums(w.dna[,names(nvr.act)]))
```

```
## [1] 35765
```

```r
sum(rowSums(w.dna[,names(yes.act)]))
```

```
## [1] 594544
```

```r
sum(rowSums(w.dna[,names(nvr.act)])) / sum(rowSums(w.dna))
```

```
## [1] 0.05674201
```

```r
# of taxa who became active, what fraction of the active community did they represent?
sum(rowSums(w.rna[,names(nvr.act)]))
```

```
## [1] 0
```

```r
sum(rowSums(w.rna[,names(yes.act)]))
```

```
## [1] 624979
```

```r
sum(rowSums(w.rna[,names(nvr.act)])) / sum(rowSums(w.rna))
```

```
## [1] 0
```

```r
sum(rowSums(w.rna[,names(yes.act)])) / sum(rowSums(w.rna))
```

```
## [1] 0.9915438
```

```r
prop.nvr.act <- rowSums(w.dna[,nvr.act]) / rowSums(w.dna)
# cbind.data.frame(design.dna, inactive = prop.nvr.act) %>%
#   ggplot(aes(x = distance, y = inactive)) +
#   geom_point() +
#   geom_line(stat = "smooth", method = "lm", formula = y ~ x, se = F) +
#   labs(x = "Reservoir transect (m)", y = "Rel. abundance of taxa\n that are never active") +
#   scale_x_reverse()
```

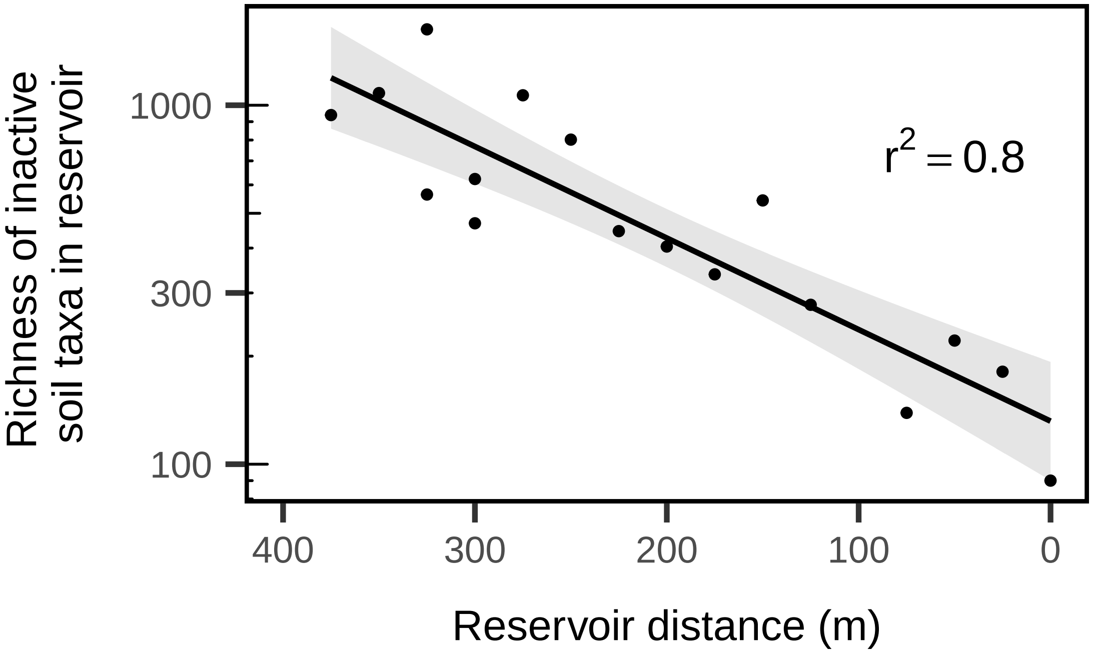
We calculate the richness of the soil taxa that are never active in the lake. We calculate richness from the DNA-based samples.

```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)	2.12	0.07745	27.37	3.215e-14
design.dna$distance	0.002551	0.0003258	7.828	1.124e-06

```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 distance (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)
```



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

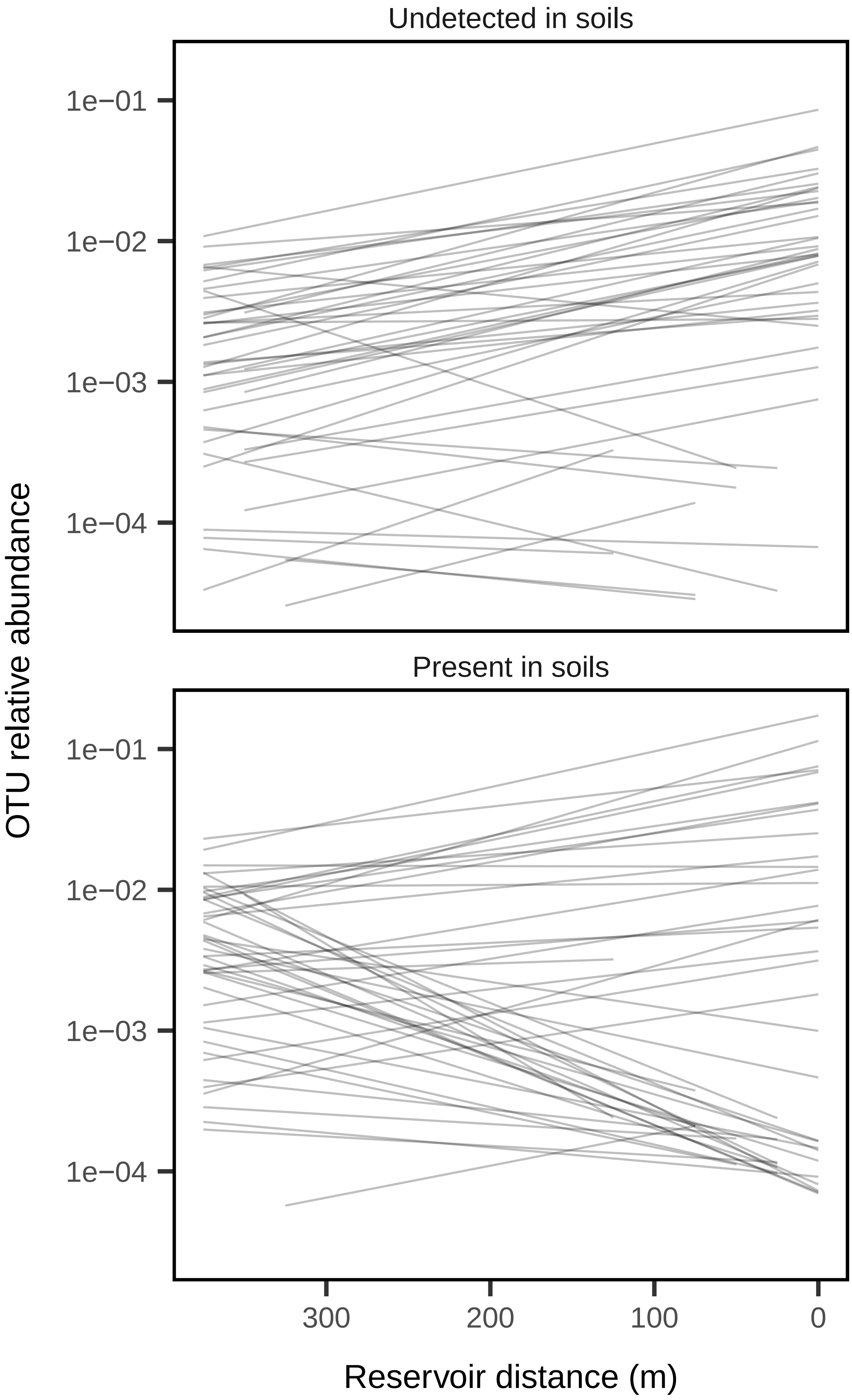
So, we observe that most soil-derived taxa appear to decay once they enter the reservoir. Do any soil-derived taxa persist in the active bacterial community of the reservoir and do they rise to high relative abundances?

```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.

```r
bind_rows(in.soil.to.plot, in.lake.to.plot) %>%
  ggplot(aes(x = distance, y = rel_abundance, group = otu_id)) +
  labs(x = "Reservoir distance (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 149 rows containing non-finite values (stat_smooth).
```


Undetected in soils
Present in soils
1e−01
1e−02
1e−03
1e−04
OTU relative abundance
300
200
100
0
Reservoir distance (m)

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

```
# 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
```

```
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
```

```
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
```

```
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

```
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
Otu00057	2.463e-05	0.03269	Bacteria	Proteobacteria

OTU	slope	pval	Domain	Phylum
Otu00138	3.152e-05	0.04589	Bacteria	Firmicutes

Table 5: Table continues below

Class	Order	Family
Gammaproteobacteria	Methylococcales	Methylococcaceae
Bacilli	Bacillales	Bacillaceae_1

Genus
Methylococcaceae_unclassified
Bacillus

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

Table 7: Core taxa not found in soils that get more common along the transect. (continued below)

OTU	slope	pval	Domain	Phylum
Otu00004	-0.0001379	3.031e-06	Bacteria	Actinobacteria
Otu00016	-5.806e-05	0.0001992	Bacteria	Actinobacteria
Otu00017	-3.298e-05	4.237e-05	Bacteria	Actinobacteria
Otu00025	-5.193e-05	0.000563	Bacteria	Actinobacteria
Otu00029	-3.389e-05	0.001212	Bacteria	Actinobacteria
Otu00031	-6.068e-05	0.000148	Bacteria	Bacteroidetes
Otu00034	-9.904e-06	2.635e-05	Bacteria	Proteobacteria
Otu00038	-4.082e-05	0.0004677	Bacteria	Actinobacteria
Otu00040	-3.681e-05	1.522e-05	Bacteria	Proteobacteria
Otu00050	-1.948e-05	0.002039	Bacteria	Bacteroidetes
Otu00055	-1.084e-05	0.006634	Bacteria	Bacteroidetes
Otu00058	-1.238e-05	0.01813	Bacteria	Armatimonadetes
Otu00071	-5.253e-05	8.694e-06	Bacteria	Planctomycetes
Otu00075	-2.21e-05	0.002713	Bacteria	Bacteria_unclassified
Otu00080	-2.261e-05	0.02962	Bacteria	Bacteroidetes
Otu00091	-1.433e-05	8.002e-05	Bacteria	Bacteroidetes
Otu00099	-2.171e-06	0.01177	Bacteria	Bacteria_unclassified
Otu00113	-1.395e-06	0.0002851	Bacteria	Bacteroidetes
Otu00118	-7.165e-06	0.01503	Bacteria	Actinobacteria
Otu00156	-9.057e-06	0.0002607	Bacteria	Bacteria_unclassified
Otu00168	-1.2e-05	0.000938	Bacteria	Bacteroidetes
Otu00178	-2.446e-06	0.02077	Bacteria	Proteobacteria

Table 8: Table continues below

Class	Order
Actinobacteria	Actinomycetales
Actinobacteria	Actinomycetales

Class	Order
Actinobacteria	Actinomycetales
Actinobacteria	Actinomycetales
Actinobacteria	Actinomycetales
Cytophagia	Cytophagales
Alphaproteobacteria	Sphingomonadales
Actinobacteria	Actinomycetales
Alphaproteobacteria	Rhodospirillales
Sphingobacteriia	Sphingobacteriales
Flavobacteriia	Flavobacteriales
Armatimonadia	Armatimonadales
Planctomycetia	Planctomycetales
Bacteria_unclassified	Bacteria_unclassified
Flavobacteriia	Flavobacteriales
Sphingobacteriia	Sphingobacteriales
Bacteria_unclassified	Bacteria_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Actinobacteria	Actinobacteria_unclassified
Bacteria_unclassified	Bacteria_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Alphaproteobacteria	Rhodobacterales

Family	Genus
Actinomycetales_unclassified	Actinomycetales_unclassified
Microbacteriaceae	Microbacteriaceae_unclassified
Actinomycetales_unclassified	Actinomycetales_unclassified
Microbacteriaceae	Microbacteriaceae_unclassified
Actinomycetales_unclassified	Actinomycetales_unclassified
Cyclobacteriaceae	Algoriphagus
Sphingomonadaceae	Sphingorhabdus
Actinomycetales_unclassified	Actinomycetales_unclassified
Acetobacteraceae	Roseomonas
Chitinophagaceae	Chitinophagaceae_unclassified
Cryomorphaceae	Cryomorphaceae_unclassified
Armatimonadaceae	Armatimonas/Armatimonadetes_gp1
Planctomycetaceae	Planctomycetaceae_unclassified
Bacteria_unclassified	Bacteria_unclassified
Flavobacteriaceae	Flavobacterium
Saprospiraceae	Saprospiraceae_unclassified
Bacteria_unclassified	Bacteria_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Actinobacteria_unclassified	Actinobacteria_unclassified
Bacteria_unclassified	Bacteria_unclassified
Bacteroidetes_unclassified	Bacteroidetes_unclassified
Rhodobacteraceae	Rhodobacteraceae_unclassified

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

Table 10: Core taxa found in soils that get rarer along the transect. (continued below)

Table 11: Table continues below

OTU	slope	pval	Domain	Phylum	Class	Order	Family	Genus
Otu00018	4.823e-05	0.02295	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas
Otu00026	1.513e-05	0.03508	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Comamonadaceae_unclassified
Otu00077	5.202e-05	0.0459	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Flavobacterium
Otu00081	2.039e-05	0.04586	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	Janthinobacterium
Otu00201	1.249e-05	0.03558	Bacteria	Acidobacteria	Acidobacteria_Gp6	Gp6	Gp6_unclassified	Gp6_unclassified
Otu00260	9.203e-06	0.0455	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Yersinia
Otu00816	2.175e-06	0.01383	Bacteria	Acidobacteria	Acidobacteria_Gp6	Gp6	Gp6_unclassified	Gp6_unclassified

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

Table 13: Core taxa found in soils that get more common along the transect. (continued below)

OTU	slope	pval	Domain	Phylum
Otu00001	-2.297e-05	0.02728	Bacteria	Proteobacteria
Otu00002	-0.000238	0.0005166	Bacteria	Actinobacteria
Otu00003	-0.0001095	0.0003038	Bacteria	Verrucomicrobia
Otu00005	-5.261e-05	0.002303	Bacteria	Bacteroidetes
Otu00006	-8.526e-06	0.04222	Bacteria	Bacteroidetes
Otu00008	-4.242e-05	0.004938	Bacteria	Actinobacteria
Otu00014	-0.000103	0.000156	Bacteria	Actinobacteria
Otu00015	-0.0001461	5.141e-05	Bacteria	Actinobacteria
Otu00096	-7.061e-06	0.006714	Bacteria	Proteobacteria
Otu00190	-3.162e-06	0.03246	Bacteria	Verrucomicrobia

Table 14: Table continues below

Class	Order	Family	Genus
Betaproteobacteria	Burkholderiales	Comamonadaceae	Comamonadaceae_unclassified
Actinobacteria	Actinomycetales	Actinomycetales_unclassified	Actinomycetales_unclassified
Spartobacteria	Spartobacteria_unclassified	Spartobacteria_unclassified	Spartobacteria_unclassified
Sphingobacteriia	Sphingobacteriales	Chitinophagaceae	Sediminibacterium
Sphingobacteriia	Sphingobacteriales	Saprospiraceae	Saprospiraceae_unclassified
Actinobacteria	Actinomycetales	Actinomycetales_unclassified	Actinomycetales_unclassified
Actinobacteria	Actinomycetales	Actinomycetales_unclassified	Actinomycetales_unclassified
Actinobacteria	Actinobacteria_unclassified	Actinobacteria_unclassified	Actinobacteria_unclassified
Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Rhodobacter
Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobiaceae	Luteolibacter

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

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

```
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
```

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

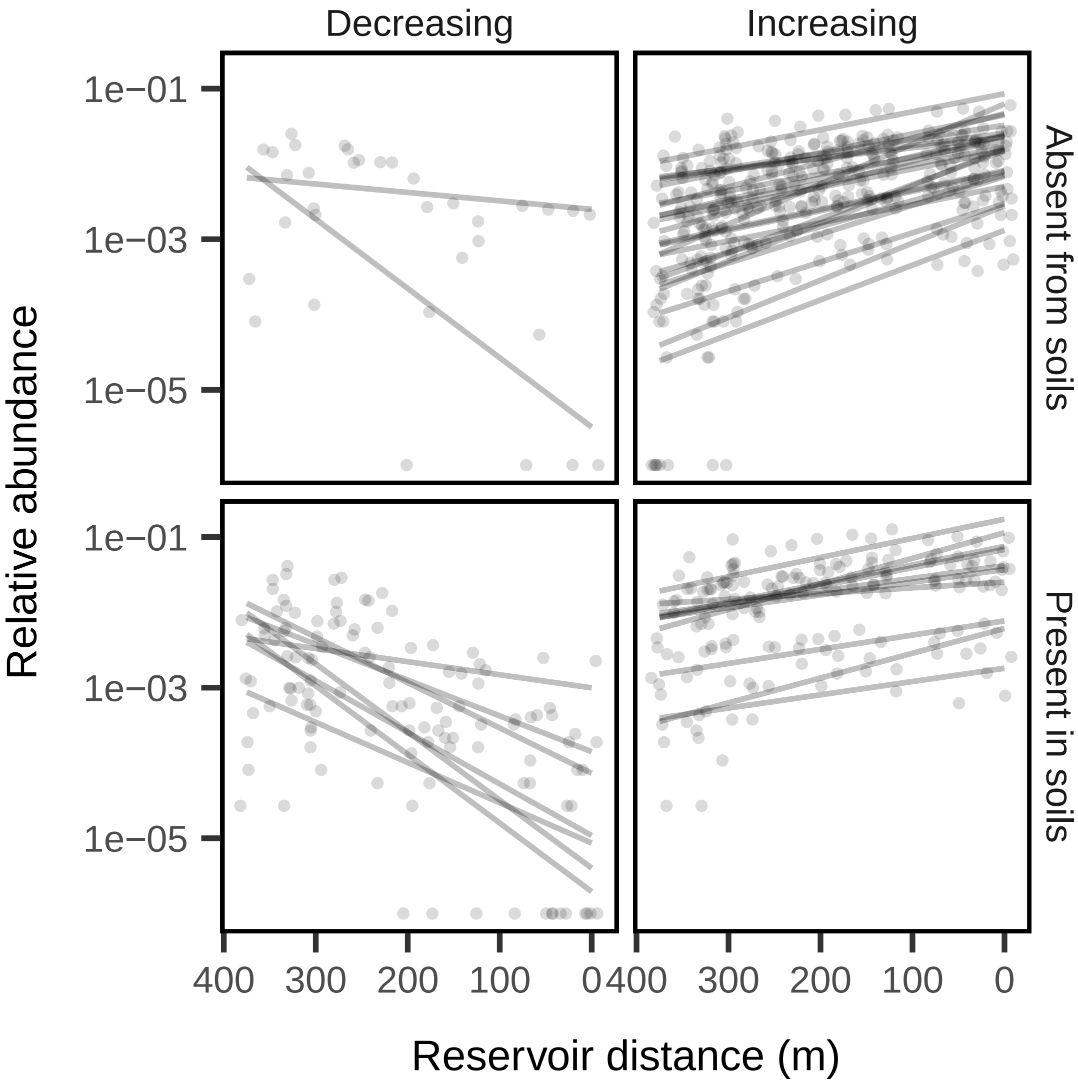
```r
df3 <- as.data.frame(OTUsREL[,soil.core.decreasing$OTU]) %>%
  rownames_to_column("sampleID") %>%
  left_join(rownames_to_column(design, "sampleID")) %>%
  gather(OTU, rel_abund, -station, -molecule, -type, -distance, -sampleID) %>%
  filter(molecule == "DNA") %>% left_join(OTU.tax) %>%
  mutate(soils = "Present in soils", change = "Decreasing")
```

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

```r
n3 <- length(unique(df3$OTU))

df4 <- as.data.frame(OTUsREL[,nonsoil.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 = "Absent from soils", change = "Decreasing")
```

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

```r
n4 <- length(unique(df4$OTU))


df.plot <- as_tibble(rbind.data.frame(df1, df2, df3, df4)) %>% 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_jitter(alpha = 0.15) +
  geom_line(stat = "smooth", alpha = 0.25, size = 1,
            color = "black", method = "lm", 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 distance (m)",
       y = "Relative abundance") +
  facet_grid(soils ~ change)
```



```
# 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))
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)

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)
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") %>%
```

```r
  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")

# 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

```r
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.pad
```

Not-included

Ecosystem functions

```r
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)
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) %
```

```r
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)
```