

Introducing R and the RStudio IDE

Alana Schick & Hena Ramay

IMC Bioinformatics, University of Calgary



UNIVERSITY OF CALGARY
CUMMING SCHOOL OF MEDICINE



**UNIVERSITY OF
CALGARY**

International
**Microbiome
Centre**



What is R?

- R (since 1995) is a programming language developed to teach statistics
- R is open source (ie. free), widely used, flexible, and powerful



Packages: the power of R

A way for the R **community** to share functions and data sets

Importing & Exporting data

From
text files, excel,
stata, SPSS , and
databases

Data Modeling

Statistical tests
Linear & non-
linear models
Machine learning
Survival analysis

Data Sharing

Plotting,
interactive plots,
reporting with
markdown and
shinny apps



What is RStudio?

RStudio is an Integrated Development Environment (IDE) that allows users to run R in a more user-friendly way



R: Engine

+



RStudio: Dashboard



Let's open Rstudio and get to know it !!

RStudio looks like this

The image displays the RStudio environment with a script for calculating alpha diversity and a boxplot of the results.

```
119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x) x / sum(x))
122
123 ## Execute filter
124 relf <- prune_taxa(keep_taxa, rel)
125 psf <- prune_taxa(keep_taxa, ps)
126
127
128
129 ##### Alpha diversity
130
131
132 ## Calculate alpha diversity using unfiltered data because rare variants influence measures of alpha div
133
134 ## Make table of alpha diversity calculations
135 alpha <- estimate_richness(psf)
136 alpha_info <- sample_data(psf)
137 aa <- cbind(alpha, alpha_info)
138
139
140 ## Check for outliers
141 aplot <- ggplot(alpha$Shannon, binwidth = 0.05) + xlab("Shannon diversity")
142 aplot <- ggplot(alpha$Simpson, binwidth = 0.005) + xlab("Simpson diversity")
143
144 ## Plot
145 a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
146 a1
147
148 a2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Simpson)") + xlab("Timepoint")
149 a2
```

The boxplot shows Alpha diversity (Shannon) on the y-axis (ranging from 2.5 to 4.0) across two timepoints: 'base' and 'week3'. For each timepoint, four treatments are compared: control (red), pre (green), pro (cyan), and syn (purple). The 'pre' treatment consistently shows the highest alpha diversity, while the 'syn' treatment shows the lowest. The 'control' and 'pro' treatments show intermediate levels of diversity. The diversity values are generally higher at the 'base' timepoint compared to 'week3'.

Timepoint	control	pre	pro	syn
base	~3.1	~3.7	~3.2	~3.4
week3	~3.3	~2.9	~3.3	~3.1

RStudio screen

The image shows a screenshot of the RStudio interface. The main window is divided into several panes:

- Script:** The left pane shows R code for calculating alpha diversity. It includes comments and functions like `transform_sample_counts`, `prune_taxa`, `estimate_richness`, and `ggplot`. A box labeled "Script" is overlaid on this pane.
- Console:** The top right pane shows the output of the R code, including the creation of OTU, sample, taxonomy, and phylogenetic trees, and the execution of the `ggplot` function. A box labeled "Console" is overlaid on this pane.
- Environment:** The bottom left pane shows the current environment with variables like `a1`, `aa`, `alpha`, `alpha_info`, `info`, `ps`, `seqtab`, `taxa`, and `tree`. A box labeled "Environment" is overlaid on this pane.
- Plots:** The bottom right pane shows a boxplot of Alpha diversity (Shannon) for two time points: "base" and "week3". Each time point has four boxplots representing different treatments: control (red), pre (green), pro (cyan), and syn (purple). A box labeled "Plots" is overlaid on this pane.

The RStudio interface includes a menu bar at the top with options like File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Window, and Help. The status bar at the bottom shows the current project as "Project: (None)" and the time as "Tue 15:43".

RStudio screen

The image shows the RStudio interface with three main panes highlighted by boxes and arrows:

- Script:** Contains R code for calculating alpha diversity. Key lines include:

```
119  
120 ## Take relative abundance  
121 rel <- transform_sample_counts(ps, function(x) x / sum(x))  
122  
123 ## Execute filter  
124 relf <- prune_taxa(keep_taxa, rel)  
125 psf <- prune_taxa(keep_taxa, ps)  
126  
127  
128  
129 ##### Alpha diversity  
130  
131  
132 ## Calculate alpha diversity using unfiltered data because rare variants influence measures of diversity  
133  
134 ## Make table of alpha diversity calculations  
135 alpha <- estimate_richness(psf)  
136 alpha_info <- sample_data(alpha)  
137 aa <- cbind(alpha, alpha_info)  
138  
139  
140 ## Check for outliers  
141 a1 <- ggplot(alpha_info, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")  
142 a2 <- ggplot(alpha_info, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Simpson)") + xlab("Timepoint")  
143  
144  
145 a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")  
146 a1  
147  
148 a2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Simpson)") + xlab("Timepoint")  
143:1 (Untitled)
```
- Console:** Shows the output of the R code, including the creation of the phyloseq object and the alpha diversity calculations.

```
> taxa_names(ps) <- asv_names  
> colnames(otu_table(ps)) <- asv_names  
> rownames(tax_table(ps)) <- asv_names  
>  
> ## Remove control samples  
> ps <- prune_samples(sample_data(ps)$treatment == "control", ps)  
> ps  
phyloseq-class experiment-level object  
otu_table() OTU Table: [ 2171 taxa and 95 samples ]  
sample_data() Sample Data: [ 95 samples by 7 sample variables ]  
tax_table() Taxonomy Table: [ 2171 taxa by 7 taxonomic ranks ]  
phy_tree() Phylogenetic Tree: [ 2171 tips and 2170 internal nodes ]  
>  
> ## Add group variable  
> sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "_"))  
> alpha <- estimate_richness(psf)  
> alpha_info <- sample_data(alpha)  
> aa <- cbind(alpha, alpha_info)  
> a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")  
> a1  
>
```
- Plots:** Displays a boxplot of Alpha diversity (Shannon) for the 'base' and 'week3' timepoints, faceted by treatment (control, pre, pro, syn). The y-axis ranges from 2.5 to 4.0. The legend indicates: control (red), pre (green), pro (cyan), and syn (purple).

Blue arrows point from the Console and Plots panes to the Script pane, indicating the flow of data from execution to the code being run.

RStudio screen

The image shows a screenshot of the RStudio software interface. The interface is divided into several panes:

- Script Pane:** Contains R code for data analysis, including functions for transforming sample counts, filtering taxa, calculating alpha diversity, and creating a phylogenetic tree. A box labeled "Script" is overlaid on this pane.
- Console Pane:** Shows the output of the R code, including the creation of a phyloseq object and the addition of a group variable. A box labeled "Console" is overlaid on this pane.
- History Pane:** Displays a list of executed R commands, such as creating a phyloseq object, making a tree, and calculating alpha diversity. A box labeled "History" is overlaid on this pane.
- Files Pane:** Shows a file browser view of the user's home directory, listing folders like .Rhistory, Applications, Desktop, Documents, Downloads, Dropbox, Library, Movies, Music, Pictures, Public, and Zotero. A box labeled "Files" is overlaid on this pane.

The top of the window shows the RStudio menu bar (File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Window, Help) and the system tray with various icons and the date/time (Tue 16:27). The bottom of the image shows the macOS dock with various application icons.

How to R – 2 ways

The image shows the RStudio interface with three main components highlighted:

- Script:** The editor window on the left contains R code for calculating alpha diversity. A box labeled "Script" is placed over the code.
- Console:** The terminal window on the right shows the execution of the R code. A box labeled "Console" is placed over the terminal output.
- Figure:** A boxplot in the Plots pane shows "Alpha diversity (Shannon)" on the y-axis (ranging from 2.5 to 4.0) and "base" and "week3" on the x-axis. The boxplot is faceted by treatment (control, pre, pro, syn). A box labeled "1. Typing and executing commands interactively here." is placed over the console and the top part of the plot.

The Environment pane at the bottom left shows the following objects:

Object	Description
a1	List of 9
aa	95 obs. of 17 variables
alpha	95 obs. of 9 variables
alpha_info	95 obs. of 8 variables
info	96 obs. of 7 variables
ps	Large phyloseq (1.5 Mb)
seqtab	Large matrix (208416 elements, 1.8 Mb)
taxa	Large matrix (15197 elements, 1.1 Mb)
tree	Large phylo (4 elements, 1 Mb)

The boxplot shows the distribution of Alpha diversity (Shannon) for each treatment group at two time points: base and week3. The y-axis ranges from 2.5 to 4.0. The legend indicates the following colors for treatments: control (red), pre (green), pro (cyan), and syn (purple). The boxplots show the median, interquartile range, and whiskers for each group. The 'pre' treatment shows the highest alpha diversity at the 'base' time point, while the 'syn' treatment shows the lowest.

How to R – 2 ways

The image shows the RStudio interface with three main components highlighted:

- Script:** The editor window on the left contains R code for calculating alpha diversity and plotting it. A yellow circle highlights the code, and a blue arrow points to the Environment pane.
- Console:** The terminal window on the right shows the execution of the R code. A box labeled "Console" and a larger box labeled "1. Typing and executing commands interactively here." are overlaid on this area.
- Environment:** The bottom-left pane shows the current environment with variables like 'a1', 'aa', 'alpha', etc. A box labeled "2. Using and saving scripts (ie. plain text files of code)." is overlaid on this area, with a blue arrow pointing to the Environment pane.
- Plots:** The bottom-right pane shows a boxplot of Alpha diversity (Shannon) for 'base' and 'week3' timepoints, categorized by treatment (control, pre, pro, syn). A box labeled "Choose this one." is overlaid on the Environment pane, pointing to the 'a1' variable.

```
119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x) x / sum(x))
122
123 ## Execute filter
124 relf <- prune_taxa(keep_taxa, rel)
125 psf <- prune_taxa(keep_taxa, ps)
126
127
128
129 ##### Alpha diversity
130
131
132 ## Calculate alpha diversity using phyloseq
133
134 ## Make table of alpha diversity
135 alpha <- estimate_richness(psf)
136 alpha_info <- sample_data(psf)
137 aa <- cbind(alpha, alpha_info)
138
139
140 ## Check for outliers
141 a1 <- ggplot(alpha$Shannon, binwidth = 0.1) + geom_boxplot(outlier.fill = NULL, outlier
142 .shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point",
143 shape = 4, size = 4, position = position_dodge(width = 0.75)) + xlab("Alpha diversity (Shannon)") + xlab
144 ("Timepoint")
145 a1
146
147
148 a2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier
149 .shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point",
150 shape = 4, size = 4, position = position_dodge(width = 0.75)) + xlab("Alpha diversity (Shannon)") + xlab
151 ("Timepoint")
152 a2
```

```
> taxa_names(psf) <- asv_names
> colnames(otu_table(psf)) <- asv_names
> rownames(tax_table(psf)) <- asv_names
>
> ## Remove control samples
> ps <- prune_samples(sample_data(psf)$treatment, ps)
> ps
phyloseq-class experimental design OTU table
otu_table() OTU
sample_data() Sample
tax_table() Taxa
phy_tree() Phylogenetic tree
>
> ## Add group variable
> sample_data(psf)$treatment <- "control"
> alpha <- estimate_richness(psf)
> alpha_info <- sample_data(psf)
> aa <- cbind(alpha, alpha_info)
> a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape
= 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size =
4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
> a1
```

Timepoint	Treatment	Alpha diversity (Shannon)
base	control	~3.4
	pre	~3.7
	pro	~3.2
	syn	~3.4
week3	control	~3.3
	pre	~2.9
	pro	~3.3
	syn	~3.1

Scripts

```
1 # jodi_btbr project, Alana Schick, April 2019
2 # This is a script to analyze the output tables of the DADA2 workflow in phyloseq
3 # Have two output files from dada2 - a sequence table and a taxonomy table, read them into R using the readRDS()
  function
4 # The formatted sample metadata is in a table called "jodi_btbr_metadata.txt"
5
6 library(phyloseq)
7 #packageVersion("phyloseq")
8 library(ggplot2)
9 #packageVersion("ggplot2")
10 library(ape)
11 library(viridis)
12 library(grid)
13 library(gridExtra)
14 library(reshape2)
15 library(DESeq2)
16 library(fields)
17 library(vegan)
18 library(ggpubr)
19 library(plyr)
20 library(RColorBrewer)
21
22 path_to_project <- "/Users/alanaschick/Drop
23
24 # Read in files
25 seqtab <- readRDS(file.path(path_to_project,
26 taxa <- readRDS(file.path(path_to_project,
27 info <- read.table(file.path(path_to_project, "jodi_btbr_metadata.txt"), header=TRUE, as.is=TRUE)
28
29 # Match sample names
30 rownames(info) <- rownames(seqtab)
31
32 # Make a phyloseq object
38:1 # (Untitled) ↕
```

Everything in the console will be forgotten when you close the session.

Scripts are saved, keeping a complete record of the commands you ran so you can run them again (ie. completely reproducible).

Can execute parts of this or the entire script.

Scripts - commenting

```
1 # jodi_btbr project, Alana Schick, April 2019
2 # This is a script to analyze the output tables of the DADA2 workflow in phyloseq
3 # Have two output files from dada2 - a sequence table and a taxonomy table, read them into R using the readRDS()
  function
4 # The formatted sample metadata is in a table called "jodi_btbr_metadata.txt"
5
6 library(phyloseq)
7 #packageVersion("phyloseq")
8 library(ggplot2)
9 #packageVersion("ggplot2")
10 library(ape)
11 library(viridis)
12 library(grid)
13 library(gridExtra)
14 library(reshape2)
15 library(DESeq2)
16 library(fields)
17 library(vegan)
18 library(ggpubr)
19 library(plyr)
20 library(RColorBrewer)
21
22 path_to_project <- "/Users/alanaschick/ Dropbox/ tmc/ projects/ jodi_btbr/"
23
24 # Read in files
25 seqtab <- readRDS(file.path(path_to_project, "results/seqtab_final.rds"))
26 taxa <- readRDS(file.path(path_to_project, "results/taxa_final.rds"))
27 info <- read.table(file.path(path_to_project, "jodi_btbr_metadata2.txt"), header = TRUE)
28
29 # Match sample names
30 rownames(info) <- rownames(seqtab)
31
32 # Make a phyloseq object
```

Comment out lines of your scripts by using the # symbol. R will not run these.

Be descriptive. You will not remember what you did a year later.

Packages

```
1 # jodi_btbr project, Alana Schick, April 2019
2 # This is a script to analyze the output tables of the DADA2 workflow in phyloseq
3 # Have two output files from dada2 - a sequence table and a taxonomy table, read them into R using the readRDS()
  function
4 # The formatted sample metadata is in a table called "jodi_btbr_metadata.txt"
5
6 library(phyloseq)
7 #packageVersion("phyloseq")
8 library(ggplot2)
9 #packageVersion("ggplot2")
10 library(ape)
11 library(viridis)
12 library(grid)
13 library(gridExtra)
14 library(reshape2)
15 library(DESeq2)
16 library(fields)
17 library(vegan)
18 library(ggpubr)
19 library(plyr)
20 library(RColorBrewer)
21
22 path_to_project <- "/Users/alanaschick/
23
24 # Read in files
25 seqtab <- readRDS(file.path(path_to_project, "results/seqtab_final.rds"))
26 taxa <- readRDS(file.path(path_to_project, "results/taxa_final.rds"))
27 info <- read.table(file.path(path_to_project, "jodi_btbr_metadata2.txt"), header = TRUE)
28
29 # Match sample names
30 rownames(info) <- rownames(seqtab)
31
32 # Make a phyloseq object
```

Packages are collections of R functions developed for a specific task.

Packages need to first be installed on your computer.

After installed, library() is the command used to load a package.

Packages

The image shows the RStudio interface with the following components:

- Code Editor:** Contains R code for data processing and alpha diversity calculations. A blue arrow points to the **Tools** menu.
- Console:** Shows the output of R commands, including the creation of a phyloseq object and the execution of a plot function.
- Environment:** Lists variables such as `a1`, `aa`, `alpha`, `alpha_info`, `info`, `ps`, `seqtab`, `taxa`, and `tree`.
- Install Packages Dialog:** A modal window with the following fields:
 - Install from:** Repository (CRAN)
 - Packages (separate multiple with space or comma):** (Empty text box)
 - Install to Library:** /Library/Frameworks/R.framework/Versions/3.6/Resources/libr
 - Install dependencies
- Boxplot:** A plot showing alpha diversity (Shannon) for different treatments (control, pre, pro, syn) across two timepoints (base and week3).

Pay close attention to the next few slides

About half of the students in every workshop have problems in understanding the concept of working directory!!

Working directory

Every time you open RStudio, it goes to a default directory, usually your home directory.

You can use the command **setwd()** to change the working directory.

```
setwd("home/aschick/projects/workshop")
```

The screenshot displays the RStudio environment. The top pane shows R code for calculating alpha diversity. The middle pane shows the Environment pane with variables like 'a1', 'aa', 'alpha', 'alpha_info', 'info', 'ps', 'seqtab', 'taxa', and 'tree'. The bottom pane shows a boxplot of Alpha diversity (Shannon) for 'base' and 'week3' timepoints, categorized by treatment (control, pre, pro, syn). The boxplot shows that the 'pre' treatment (green) has the highest alpha diversity at both timepoints, while the 'syn' treatment (purple) has the lowest. The 'control' (red) and 'pro' (cyan) treatments show intermediate levels of diversity.

```
119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x)
122
123 ## Execute filter
124 relf <- prune_taxa(keep_taxa, rel)
125 psf <- prune_taxa(keep_taxa, ps)
126
127
128
129 ##### Alpha diversity
130
131
132 ## Calculate alpha diversity using unfiltered d
133
134 ## Make table of alpha diversity calculations
135 alpha <- estimate_richness(psf)
136 alpha_info <- sample_data(psf)
137 aa <- cbind(alpha, alpha_info)
138
139
140 ## Check for outliers
141 aplot(alpha$Shannon, binwidth = 0.05) + xlab("S
142 aplot(alpha$Simpson, binwidth = 0.005) + xlab("Simpson diversity")
143
144 ## Plot
145 a1 <- ggplot(aa, aes(x = "timepoint", y = Shannon)) +
146   .shape = 21) + scale_fill
147   shape = 4, size = 4, posi
148   ("Timepoint")
149 a1
150
151 a2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape
152   R Script >
153:1 (Untitled) >
```

Environment

Variable	Description
a1	List of 9
aa	95 obs. of 17 variables
alpha	95 obs. of 9 variables
alpha_info	95 obs. of 8 variables
info	96 obs. of 7 variables
ps	Large phyloseq (1.5 Mb)
seqtab	Large matrix (208416 elements, 1.8 Mb)
taxa	Large matrix (15197 elements, 1.1 Mb)
tree	Large phylo (4 elements, 1 Mb)

Alpha diversity (Shannon)

base week3

treatment

- control
- pre
- pro
- syn

Relative paths


▼  Users

▼  hena

▼  projects

▼  rworkshop

▶  data

 X.csv

▶  analysis

← Setwd("Users/hena/projects/rworkshop/")

read.csv("data/x.csv")

```
119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x) x / sum(x))
122
123 ## Execute filter
124 relf <- prune_taxa(keep_taxa, rel)
125 psf <- prune_taxa(keep_taxa, ps)
126
127
128
129 ##### Alpha diversity
130
131
132 ## Calculate alpha diversity using unfiltered data because rare variants influence measures of alpha div
133
134 ## Make table of alpha diversity calculations
135 alpha <- estimate_richness(ps)
136 alpha_info <- sample_data(ps)
137 aa <- cbind(alpha, alpha_info)
138
```

```
138:1 (Untitled) R Script
Environment History Connections
# Match sample names
rownames(info) <- rownames(seqtab)
# Make a phyloseq object
ps <- phyloseq(otu_table(seqtab, taxa_are_rows=FALSE), sample_data(info), tax_table(taxa))
## Make a tree and add the tree to a new phyloseq object
tree <- rtree(ntaxa(ps), rooted = TRUE, tip.label = taxa_names(ps))
ps <- phyloseq(otu_table(seqtab, taxa_are_rows=FALSE), sample_data(info), tax_table(taxa), phy_tree(tree))
asv_names <- vector(dim(otu_table(ps))[2], mode = "character")
for (i in 1:dim(otu_table(ps))[2]){
  asv_names[i] <- paste("ASV", i, sep = "_")
}
taxa_names(ps) <- asv_names
colnames(otu_table(ps)) <- asv_names
rownames(tax_table(ps)) <- asv_names
## Remove control samples
ps <- prune_samples(sample_data(ps)$treatment != "NA", ps)
ps
## Add group variable
sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "_"))
alpha <- estimate_richness(ps)
alpha_info <- sample_data(ps)
aa <- cbind(alpha, alpha_info)
a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape =
a1
```

```
Console Terminal R Markdown
~/
> taxa_names(ps) <- asv_names
> colnames(otu_table(ps)) <- asv_names
> rownames(tax_table(ps)) <- asv_names
>
> ## Remove control samples
> ps <- prune_samples(sample_data(ps)$treatment != "NA", ps)
> ps
phyloseq-class experiment-level object
otu_table() OTU Table: [ 2171 taxa and 95 samples ]
sample_data() Sample Data: [ 95 samples by 7 sample variables ]
tax_table() Taxonomy Table: [ 2171 taxa by 7 taxonomic ranks ]
phy_tree() Phylogenetic Tree: [ 2171 tips and 2170 internal nodes ]
>
> ## Add group variable
> sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "_"))
> alpha <- estimate_richness(ps)
> alpha_info <- sample_data(ps)
> aa <- cbind(alpha, alpha_info)
> a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape
= 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size =
4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
> a1
>
```

Files Plots Packages Help Viewer

- New Folder
- Delete
- Rename
- More

Name	Size	Modified
.Rhistory	1.1 KB	Feb 11, 2019, 5:41 PM

The directory that you see here is not necessarily the same as your working directory. Please do not use this to find your files.

Screenshot

Working directory

The screenshot displays the RStudio interface. The main editor shows R code for calculating alpha diversity. The Environment pane on the left lists variables: a1 (List of 9), aa (95 obs. of 17 variables), alpha (95 obs. of 9 variables), alpha_info (95 obs. of 8 variables), info (96 obs. of 7 variables), ps (Large phyloseq (1.5 Mb)), seqtab (Large matrix (208416 elements, 1.8 Mb)), taxa (Large matrix (15197 elements, 1.1 Mb)), and tree (Large phylo (4 elements, 1 Mb)). The bottom right pane shows a boxplot of Alpha diversity (Shannon) for 'base' and 'week3' timepoints, with treatments 'control', 'pre', 'pro', and 'syn' represented by different colors. The y-axis ranges from 2.5 to 4.0. The boxplots show the distribution of alpha diversity for each treatment at each timepoint, with individual data points overlaid as 'point' geom.

```
119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x)
122
123 ## Execute filter
124 relf <- prune_taxa(keep_taxa, rel)
125 psf <- prune_taxa(keep_taxa, ps)
126
127
128
129 ##### Alpha diversity
130
131
132 ## Calculate alpha diversity using unfiltered data
133
134 ## Make table of alpha diversity calculations
135 alpha <- estimate_richness(psf)
136 alpha_info <- sample_data(psf)
137 aa <- cbind(alpha, alpha_info)
138
139
140 ## Check for outliers
141 a1 <- ggplot(alpha$Shannon, binwidth = 0.05) + xlab("Timepoint")
142 a2 <- ggplot(alpha$Simpson, binwidth = 0.005) + xlab("Timepoint")
143
144 ## Plot
145 a1 <- ggplot(aa, aes(x = timepoint, y = Shannon)) +
146   .shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point",
147   shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
148 a2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) +
149   .shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point",
150   shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Simpson)") + xlab("Timepoint")
151 a1
152 a2
```

Environment

Variable	Description
a1	List of 9
aa	95 obs. of 17 variables
alpha	95 obs. of 9 variables
alpha_info	95 obs. of 8 variables
info	96 obs. of 7 variables
ps	Large phyloseq (1.5 Mb)
seqtab	Large matrix (208416 elements, 1.8 Mb)
taxa	Large matrix (15197 elements, 1.1 Mb)
tree	Large phylo (4 elements, 1 Mb)

Alpha diversity (Shannon)

base week3

control pre pro syn

However: you may want to run your script on a different computer with a different directory structure where that directory does not exist.

Or you may want to work in multiple directories.

RStudio Project

The screenshot shows the RStudio interface with a code editor on the left containing R code for calculating alpha diversity. A blue arrow points to the 'File' menu. A box highlights the text 'File > New Project...'. A 'New Project' dialog box is open in the center, showing three options: 'New Directory', 'Existing Directory', and 'Version Control'. In the background, a boxplot shows 'Alpha div' on the y-axis (ranging from 2.5 to 3.0) and 'base' and 'week3' on the x-axis. The boxplot is faceted by 'treatment' (control, pre, pro, syn). A 'Screenshot' button is visible at the bottom of the boxplot area.

File > New Project...

Clicking on New Directory will create an RStudio Project.

This directory will have all the data, files, plots, etc. for that project as well as a .Rproj file.

Error messages

```
Console Terminal x R Markdown x
~/ ↵
>
> ## Remove control samples
> ps <- prune_samples(sample_data(ps)$treatment != "NA", ps)
> ps
phyloseq-class experiment-level object
otu_table() OTU Table: [ 2171 taxa and 95 samples ]
sample_data() Sample Data: [ 95 samples by 7 sample variables ]
tax_table() Taxonomy Table: [ 2171 taxa by 7 taxonomic ranks ]
phy_tree() Phylogenetic Tree: [ 2171 tips and 2170 internal nodes ]
>
> ## Add group variable
> sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "_"))
> alpha <- estimate_richness(ps)
> alpha_info <- sample_data(ps)
> aa <- cbind(alpha, alpha_info)
> a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
> a1
> ord1 <- ordinate(relf, method = "NMDS", distance = "bray")
Error in ordinate(relf, method = "NMDS", distance = "bray") :
  object 'relf' not found
> b1 <- plot_ordination(relf, ord1, color = "timepoint", shape = "treatment", title = "NMDS - Bray") + scale_colour_manual(values = viridis(3))
Error in plot_ordination(relf, ord1, color = "timepoint", shape = "treatment", :
  object 'relf' not found
> b1
Error: object 'b1' not found
> |
```

Error messages

```
Console Terminal x R Markdown x
~/ ↵
>
> ## Remove control samples
> ps <- prune_samples(sample_data(ps)$tr
> ps
phyloseq-class experiment-level object
otu_table() OTU Table: [ 2171
sample_data() Sample Data: [ 95 sa
tax_table() Taxonomy Table: [ 2171
phy_tree() Phylogenetic Tree: [ 2171
>
> ## Add group variable
> sample_data(ps)$group <- factor(paste(
> alpha <- estimate_richness(ps)
> alpha_info <- sample_data(ps)
> aa <- cbind(alpha, alpha_info)
> a1 <- ggplot(aa, aes(x = timepoint, y
= 21) + scale_fill_manual(values = rainb
4, position = position_dodge(width = 0.7
> a1
> ord1 <- ordinate(relf, method = "NMDS"
Error in ordinate(relf, method = "NMDS",
object 'relf' not found
> b1 <- plot_ordination(relf, ord1, colo
ual(values = viridis(3))
Error in plot_ordination(relf, ord1, col
object 'relf' not found
> b1
Error: object 'b1' not found
> |
```



```
tment, sep = "_"))
```

```
tlier.fill = NULL, outlier.shape  
om = "point", shape = 4, size =  
"Timepoint")
```

```
NMDS - Bray") + scale_colour_man
```

Getting help

1) Search in Help tab

2) Type ? followed by the function name in the console (or ?? for installed packages)

```
119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x) x / sum(x))
122
123 ## Execute filter
124 relf <- prune_taxa(keep_taxa, rel)
125 psf <- prune_taxa(keep_taxa, ps)
126
127
128
129 ##### Alpha diversity
130
131
132 ## Calculate alpha diversity using unfiltered data because rare v
133
134 ## Make table of alpha diversity calculations
135 alpha <- estimate_richness(psf)
136 alpha_info <- sample_data(psf)
137 aa <- cbind(alpha, alpha_info)
138
139
140 ## Check for outliers
141 aplot(alpha$Shannon, binwidth = 0.05) + xlab("Shannon diversity")
142 aplot(alpha$Simpson, binwidth = 0.005) + xlab("Simpson diversity")
143
144 ## Plot
145 a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment))
146   .shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + s
147   shape = 4, size = 4, position = position_dodge(width = 0.75)) + y
148   ("Timepoint")
149 a1
150
151 a2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier
152   R Script >
153:1 (Untitled) >
```

Environment

Object	Description
a1	List of 9
aa	95 obs. of 17 variables
alpha	95 obs. of 9 variables
alpha_info	95 obs. of 8 variables
info	96 obs. of 7 variables
ps	Large phyloseq (1.5 Mb)
seqtab	Large matrix (208416 elements, 1.8 Mb)
taxa	Large matrix (15197 elements, 1.1 Mb)
tree	Large phylo (4 elements, 1 Mb)

Values

Legend:

- control
- pre
- pro
- syn

Boxplot showing Alp (Y-axis, 2.5 to 4.0) vs timepoint (X-axis: base, week3). The plot shows four boxplots for each timepoint, colored by treatment: control (red), pre (green), pro (cyan), and syn (purple). Outliers are marked with 'x'.

Console:

```
> taxa_names(psf) <- asv_names
> colnames(otu_table(psf)) <- asv_names
```


Getting help

- 1) Search in Help tab
- 2) Type ? followed by the function name in the console (or ?? for installed packages)
- 3) Google the error message



See website for tips and resources!

The internet will make those bad words go away



Essential

Googling the
Error Message

O RLY?

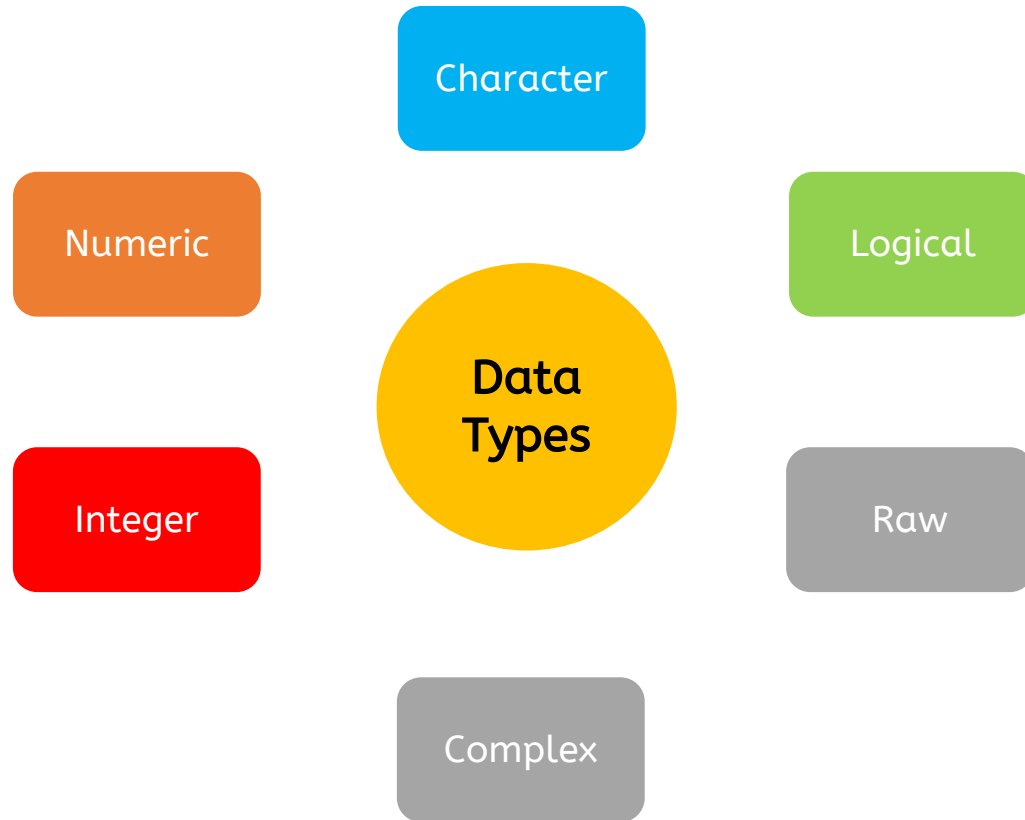
The Practical Developer
@ThePracticalDev

Summary and best practices

- Always save your code in R scripts
- Load packages using `library()` at the top of your script
- Write clear, readable code with comments*
- Be mindful of your working directory or location of files
- Use RStudio projects to organize scripts, data, and output

*See <http://adv-r.had.co.nz/Style.html> for tips.

Data Types



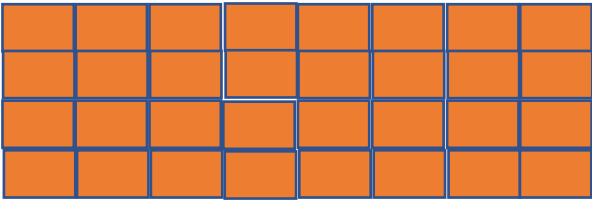
Data Structures

Vector
1d, homogeneous



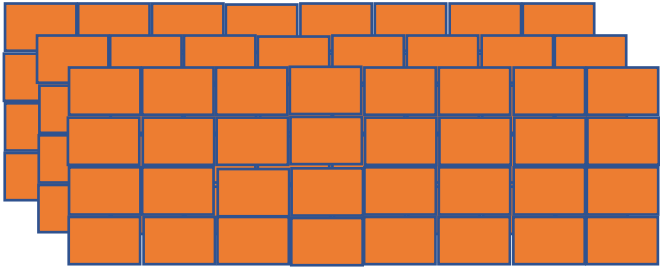
Matrix

2d, homogeneous



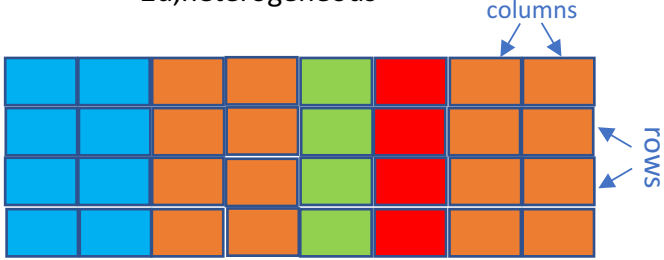
Array

3d, homogeneous



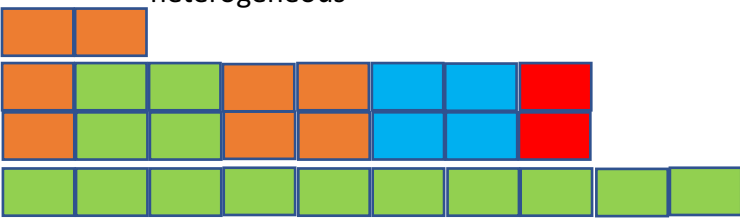
Dataframe & tibble

2d, heterogeneous



Lists

heterogeneous



Homogeneous means that it can hold only one data type at a time.
Heterogeneous means it can hold multiple datatypes at a time.

Data Structures

Vector

1d, homogeneous

45	56	62	92	23	39	67	84
----	----	----	----	----	----	----	----

Dataframe & tibble

2d, heterogeneous

a	cat	23					
b	dog	34					
c	bat	5					
d	bee	0.4					

columns

rows

Data Structures

Vector

1d, homogeneous

Index	1	2	3	4	5	6	7	8
	45	56	62	92	23	39	67	84

Dataframe & tibble

2d, heterogeneous

Index = row,col	1,1	1,2	1,3					
1,1	a	cat	23					
2,1	b	dog	34					
3,1	c	bat	5					
	d	bee	0.4					

columns

rows