Introducing R and the RStudio IDE

Alana Schick, IMC Bioinformatics, University of Calgary

a.schick@ucalgary.ca, @alana_schick

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 for everything

car – car's Anova function is popular for making type II and type III Anova tables

mgcv – Generalized Additive Models

lme4/nlme – Linear and non-linear mixed effects models

randomForest – Random forest methods from machine learning

multcomp – Tools for multiple comparison testing

vcd – Visualization tools and tests for categorical data

glmnet – Lasso and elastic-net regression methods with cross validation

survival – Tools for survival analysis

caret – Tools for training regression and classification models

https://support.rstudio.com/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages

Packages for everything

```
> install.packages("meme")
> library(meme)
> meme(img, "data science",
    "so hot right now")
```

What is R? What is RStudio?

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

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

What is R? What is RStudio?

R: Engine	RStudio: Dashboard

http://moderndive.com/

RStudio looks like this

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help

RStudio

index.Rmd | Schedule.Rmd* | Session1.Rmd* | jodi_btbr_phyloseq.R

Source on Save | Run | Source

```
## Take relative abundance
rel <- transform_sample_counts(ps, function(x) x / sum(x))

## Execute filter
relf <- prune_taxa(keeptaxa, rel)
psf <- prune_taxa(keeptaxa, ps)



####### Alpha diversity


## Calculate alpha diversity using unfiltered data because rare variants influence measures of alpha div

## Make table of alpha diversity calculations
alpha <- estimate_richness(ps)
alpha_info <- sample_data(ps)
aa <- cbind(alpha, alpha_info)


## Check for outliers
qplot(alpha$Shannon, binwidth = 0.05) + xlab("Shannon diversity")
qplot(alpha$Simpson, binwidth = 0.005) + xlab("Simpson diversity")

## Plot
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

a2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier
```

143:1 (Untitled) R Script

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
>
```

Environment | History | Connections

Import Dataset | List

Global Environment

Data	
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

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish

Screenshot

RStudio screen

RStudio screen

RStudio screen

How to R – 2 ways

How to R – 2 ways

Scripts

```
# jodi_btbr project, Alana Schick, April 2019
# This is a script to analyze the output tables of the DADA2 workflow in phyloseq
# Have two output files from dada2 - a sequence table and a taxonomy table, read them into R using the readRDS() function
# The formatted sample metadata is in a table called "jodi_btbr_metadata.txt"

library(phyloseq)
#packageVersion("phyloseq")
library(ggplot2)
#packageVersion("ggplot2")
library(ape)
library(viridis)
library(grid)
library(gridExtra)
library(reshape2)
library(DESeq2)
library(fields)
library(vegan)
library(ggpubr)
library(plyr)
library(RColorBrewer)

path_to_project <- "/Users/alanaschick/Drop

# Read in files
seqtab <- readRDS(file.path(path_to_projec
taxa <- readRDS(file.path(path_to_project,
info <- read.table(file.path(path_to_projec

# Match sample names
rownames(info) <- rownames(seqtab)

# Make a phyloseq object
```

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.

38:1 (Untitled) R Script

Packages

```
# jodi_btbr project, Alana Schick, April 2019
# This is a script to analyze the output tables of the DADA2 workflow in phyloseq
# Have two output files from dada2 - a sequence table and a taxonomy table, read them into R using the readRDS() function
# The formatted sample metadata is in a table called "jodi_btbr_metadata.txt"

library(phyloseq)
#packageVersion("phyloseq")
library(ggplot2)
#packageVersion("ggplot2")
library(ape)
library(viridis)
library(grid)
library(gridExtra)
library(reshape2)
library(DESeq2)
library(fields)
library(vegan)
library(ggpubr)
library(plyr)
library(RColorBrewer)

path_to_project <- "/Users/alanaschick/

# Read in files
seqtab <- readRDS(file.path(path_to_project, "results/seqtab_final.rds"))
taxa <- readRDS(file.path(path_to_project, "results/taxa_final.rds"))
info <- read.table(file.path(path_to_project, "jodi_btbr_metadata2.txt"), header = TRUE)

# Match sample names
rownames(info) <- rownames(seqtab)

# Make a phyloseq object
```

38:1 # (Untitled) R Script

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

Tools > Install packages...
Install Packages
Install from:
Configuring Repositories
Repository (CRAN)
Packages (separate multiple with space or comma):
Install to Library:
/Library/Frameworks/R.framework/Versions/3.6/Resources/libr
Install dependencies
Install
Cancel

Scripts - commenting

```
# jodi_btbr project, Alana Schick, April 2019
# This is a script to analyze the output tables of the DADA2 workflow in phyloseq
# Have two output files from dada2 - a sequence table and a taxonomy table, read them into R using the readRDS()
function
# The formatted sample metadata is in a table called "jodi_btbr_metadata.txt"

library(phyloseq)
#packageVersion("phyloseq")
library(ggplot2)
#packageVersion("ggplot2")
library(ape)
library(viridis)
library(grid)
library(gridExtra)
library(reshape2)
library(DESeq2)
library(fields)
library(vegan)
library(ggpubr)
library(plyr)
library(RColorBrewer)

path_to_project <- "/Users/alanaschick/Dropbox/IMC/projects/jodi_btbr/

# Read in files
seqtab <- readRDS(file.path(path_to_project, "results/seqtab_final.rds"))
taxa <- readRDS(file.path(path_to_project, "results/taxa_final.rds"))
info <- read.table(file.path(path_to_project, "jodi_btbr_metadata2.txt"), header = TRUE)

# Match sample names
rownames(info) <- rownames(seqtab)

# Make a phyloseq object
```

:38:1 # (Untitled) R Script

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.

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")
```

Working directory

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

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 .Rprog file.

Error messages

Console | Terminal | R Markdown

```
>
> ## 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_man
ual(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 | R Markdown

```
>
> ## 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

```
## Take relative abundance
rel <- transform_sample_counts(ps, function(x) x / sum(x))

## Execute filter
relf <- prune_taxa(keeptaxa, rel)
psf <- prune_taxa(keeptaxa, ps)



####### Alpha diversity


## Calculate alpha diversity using unfiltered data because rare variants influence measures of alpha div

## Make table of alpha diversity calculations
alpha <- estimate_richness(ps)
alpha_info <- sample_data(ps)
aa <- cbind(alpha, alpha_info)


## Check for outliers
qplot(alpha$Shannon, binwidth = 0.05) + xlab("Shannon diversity")
qplot(alpha$Simpson, binwidth = 0.005) + xlab("Simpson diversity")

## Plot
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
```

```
> taxa_names(ps) <- asv_names
> colnames(otu_table(ps)) <- asv_names
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 ]
> ps
> ... group variable
> ... (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
>
```

Getting help

1) Search in Help tab

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

```
> ?barplot
> ??geom_point
```

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 pre-workshop lesson for tips and resources!

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.