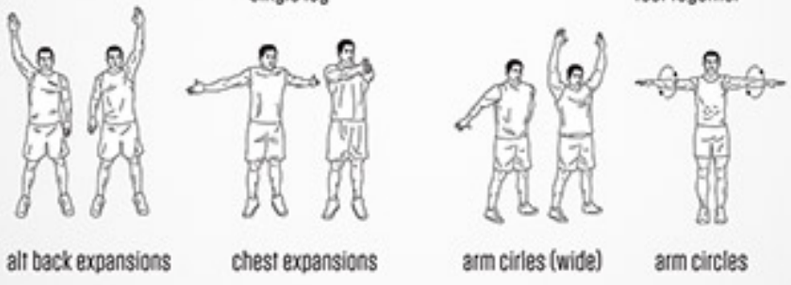
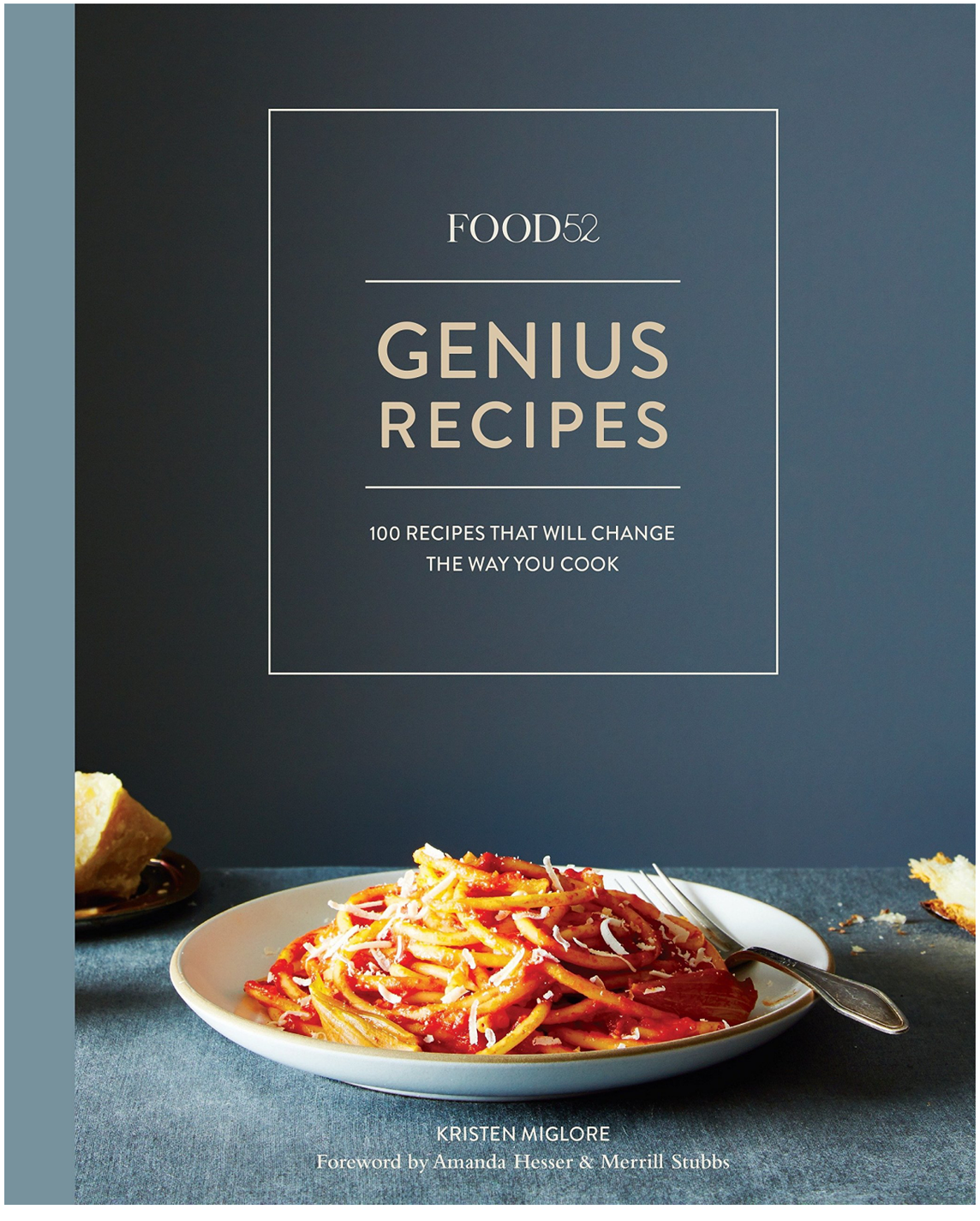
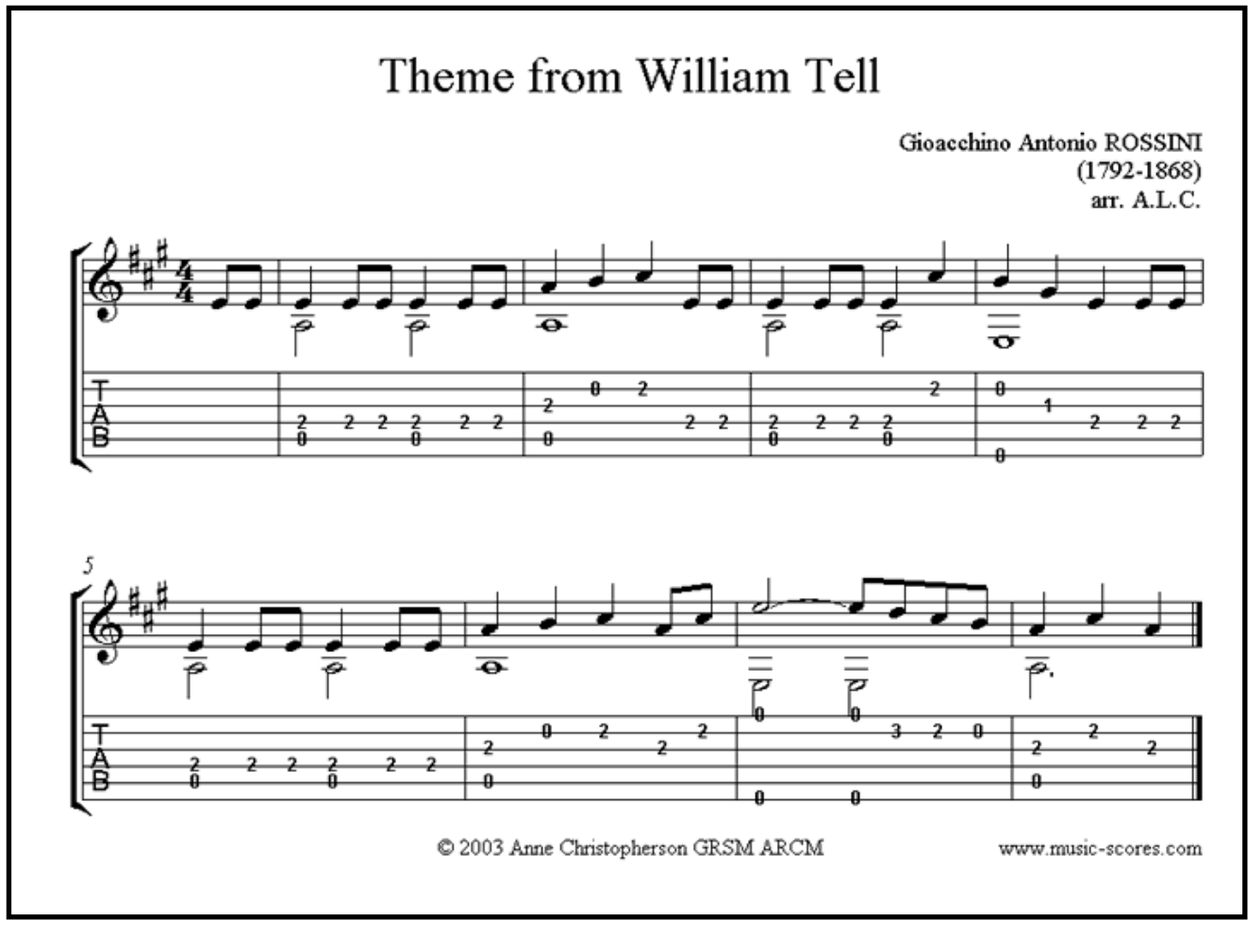
45 minutes

warmup



What is repeatability, **reproducibility**, and replicability in science and why is it important?


FOOD52
GENIUS RECIPES
100 RECIPES THAT WILL CHANGE THE WAY YOU COOK
KRISTEN MIGLORE
Foreword by Amanda Hesser & Merrill Stubbs


Theme from William Tell
Gioacchino Antonio ROSSINI
(1792-1868)
arr. A.L.C.
© 2003 Anne Christopherson GRSM ARCM
www.music-scores.com

Genetic Diversity Centre Zurich

Repeatability is a measure of the likelihood that, having produced one result from an experiment, you can try the same experiment, with the same setup, and produce that same result. It is a way for researchers to verify that their own results are true and are not just **chance artefacts**.

The **reproducibility** of data is a measure of whether a different research team can attain results published in a paper using the same methods. This shows that the results are **not artefacts of the unique setup in one research lab**. It is easy to see why reproducibility is desirable, as it reinforces findings and protects against rare cases of fraud, or less rare cases of human error, in the production of significant results.

Replicability - Different team, different experimental setup. If an observation is replicable it should be able to be made by a different team, using a different measuring system and dataset, in a different location, on multiple trials. This would therefore involve collecting data anew.

Source: https://www.technologynetworks.com/informatics/articles/repeatability-vs-reproducibility

"Science is not about making predictions or performing experiments. Science is about explaining."

Bill Gaede

Genetic Diversity Centre Zurich

What do you need to make your research **reproducible**?

- Take Ms&Ms serious and be precise. Use e.g. markdown reports to document your analysis.
- Avoid applications with GUIs.
- Understand the (limits of the) application your are using.
- Learn and use a code style.
- Document your scripts.
- Consider to publish your polished script(s) in supplementary (e.g. RMarkdown).
- Use services like Dryad for data files associated with your publication.


R

How would you improve the following R code snippet?

```
MV<-get_manifests(Data,blocks)¬
check_MV<-test_manifest_scaling(MV,specs$scaling)¬
gens<-get_generals(MV,path_matrix)¬
names(blocks)<-gens$lvs_names¬
block_sizes<-lengths(blocks)¬
blockinds<-indexify(blocks)¬
```

GDC Genetic Diversity Centre Zurich

How would you improve the following R code snippet?

Use a coding style, work with spaces and add comments.

```
1  # ============================================
2  # Preparing data and blocks indexification
3  # ============================================
4
5  # building data matrix 'MV'
6  MV        <- get_manifests(Data, blocks)
7  check_MV <- test_manifest_scaling(MV, specs$scaling)
8
9  # generals about obs, mvs, lvs
10 gens <- get_generals(MV, path_matrix)
11
12 # indexing blocks
13 names(blocks) <- gens$lvs_names
14 block_sizes   <- lengths(blocks)
15 blockinds     <- indexify(blocks)
```

add comments (#)

use white space

vertical alignment

```r
CalculateSampleCovariance <- function(x, y, verbose = TRUE) {
  # Computes the sample covariance between two vectors.
  #
  # Args:
  #   x: One of two vectors whose sample covariance is to be calculated.
  #   y: The other vector. x and y must have the same length, greater than one,
  #      with no missing values.
  #   verbose: If TRUE, prints sample covariance; if not, not. Default is TRUE.
  #
  # Returns:
  #   The sample covariance between x and y.
  n <- length(x)
  # Error handling
  if (n <= 1 || n != length(y)) {
    stop("Arguments x and y have different lengths: ",
         length(x), " and ", length(y), ".")
  }
  if (TRUE %in% is.na(x) || TRUE %in% is.na(y)) {
    stop(" Arguments x and y must not have missing values.")
  }
  covariance <- var(x, y)
  if (verbose)
    cat("Covariance = ", round(covariance, 4), ".\n", sep = "")
  return(covariance)
}
```

Loading R packages: What is the difference between the following R commands?

```
(A) install.packages("dplyr")
(B) load("dplyr")
(C) source("dplyr")
(D) require("dplyr")
(E) library("dplyr")
```

Loading R packages: What is the difference between the following R commands?

```
(A) install.packages("dplyr")
(B) load("dplyr")
(C) source("dplyr")
(D) require("dplyr")
(E) library("dplyr")
```

GDC Genetic Diversity Centre Zurich

install.packages

```
# Install R packages from repositories (or local files)
install.packages("dplyr")
install.packages(c("dplyr", "plyr")
install.packages("lattice", repos="http://cran.r-project.org")
install.packages("local.R.Package", lib.loc="/my/local/R/library")
```

save & load

```
## Reload datasets written with the function save

# save object x
save(x, file = "test.rda")
# save all objects
save(list = ls(all = TRUE), file= "all.rda")
# load dataset(s)
load("test.rda", envir = parent.frame(), verbose = FALSE)
```

source

```
## It is good practice to create separate R scripts that
## you can use to store sets of related functions.
## You can then call those functions using the source() function,
## at the top of your script. R will then load those functions into
## memory and you can use them!
source("my_awesome_functions.R")
```

library & require

The library() and require() can be used to attach and load add-on packages which are already installed.

The **library**() by default returns an error if the requested package does not exist.

The **require**() is designed to be used inside functions as it gives a warning message and returns a logical value say, FALSE if the requested package is not found and TRUE if the package is loaded.

It is better to use the library() as it gives the error message if the package is not found during the package loading time. This will indeed avoid unnecessary headaches of tracking down the errors caused while attempting to use the library routines which are not installed.

What is the first and should be the last line of code in a new R session/script?

GDC Genetic Diversity Centre Zurich

First and last line of R code

First

```
## clean/reset environment
rm(list = ls())
```

Last

```
## Session-Log
session.log <- sessionInfo()
write("session.log")
```

What is the difference between the two R commands (A) and (B)?

```
(A) brewer.pal.info["Blues",]
(B) RColorBrewer::brewer.pal.info["Blues",]
```

What is the difference between the two R commands?

```
# Use a particular function from a specific package
RColorBrewer::brewer.pal.info["Blues",]
```

```
# The function only works if the package is installed and loaded
library("Blues")
brewer.pal.info["Blues",]
# The function works without loading the packege
RColorBrewer::brewer.pal.info["Blues",]
```

What is the problem with the following R code snippet:

```
a <- 12
b <- 23
c <- 45
z <- (a * b) / c
```

What is the problem with the following R code snippet:

```
a <- 12
b <- 23
c <- 45
z <- (a * b) / c
Y <- c(a,b)
```

c {base} - combine values into a vector or list

What is the problem with the following R code snippet:

```
# A simple test to check
# for build-in functions
?c
?pi
?mean
```

What can I do to make random objects reproducable in R?

```
# Generate random number(s)
> rnorm(1)
# -1.74678
> rnorm(1)
# 1.108086
```

What can I do to make random objects reproducible in R?

```
# Generate random number(s)
> set.seed(190815); rnorm(1)
# 1.564598
> set.seed(190815); rnorm(1)
# 1.564598
```

What is the difference between R code for version A and B?

```
x <- matrix( rnorm( 50000 * 900),
        nrow = 50000,
        ncol =   900)
t(x) %*% x    # version A
crossprod(x) # version B
```

What is the difference between R code line version A and B?

```
> system.time(t(x) %*% x)
   user  system elapsed
 27.053   0.148  27.277
> system.time(crossprod(x))
   user  system elapsed
 21.166   0.020  21.209
```

Microbiota Projects

What is the difference between metagenome and amplicon (marker based) sequencing?

What is the difference between metagenome and amplicon (marker based) sequencing?

Metagenomics - You have fragments of different species and different regions of the genome (with a low(er) coverage). Careful with possible contamination.

Amplicon-Sequencing - You have data from one region of the genome of many different species. Careful with possible PCR artefacts.

What is the first thing to do once you received your sequencing data?

Genetic Diversity Centre Zurich

What is the first thing to do once you received your sequencing data?

- Quality control (see check list)
- Archiving (safe storage) raw data
- Submit raw data to data archive

Genetic Diversity Centre Zurich

What is/are the **most important** aspect/s of a microbiota project?

- ☐ Number of reads (coverage) per samples
- ☐ Number of replicates
- ☐ Negative and positive samples
- ☐ Something else …

Genetic Diversity Centre Zurich

What is/are the most important aspect/s of a microbiota project?

The project aim should shapes the project design. Thus all projects should start with a biological relevant and clear question.

Genetic Diversity Centre Zurich

What is better, (a) to apply very stringend parameters and remove bad data early or (b) use relaxed parameters and filter later?

- ☐ The earlier the better
- ☐ Later is better
- ☐ Depends
- ☐ I would use a combination

What is better, (a) to apply very stringend parameters and remove bad data early or (b) use relaxed parameters and filter later?

I am not sure since it depens very much on your dataset, the quality and your expectations.

Can I compare data from different projects (meta analysis)?

Can I compare OTUs from different projects?

GDC Genetic Diversity Centre Zurich

Can I compare data from different studies (meta analysis)?

This might be very problematic for various reasons (e.g. extraction method, primer).

Can I compare OTUs from different projects?

You have to be careful and maybe not compare OTUs (based on the same primer) just on taxonomic levels.

How many OTUs are normal?

- ☐ < 100 OTUs
- ☐ < 1,000 OTUs
- ☐ < 10,000 OTUs
- ☐ Something else …

How many OTUs are normal?

There is no normal. The number of OTUs depends very much on your system, sample and data preparation.

What method should I use do calculate alpha diversity?

- ☐ Observed / Abundance
- ☐ Chao1
- ☐ ACE
- ☐ Shannon
- ☐ Simpson
- ☐ InvSimpson
- ☐ All together …

What method should I use do calculate alpha diversity?

It is important you try different alpha diversity measures and compare the results.

Can I use alpha diversity to compare samples?

- ☐ Yes, of corse!
- ☐ No, of corse not!
- ☐ Maybe? Depense!

Can I use alpha diversity to compare samples?

Why not. You have to explore your data first and consider a few processing steps (e.g. rarefy your samples) before a comparison would make sense.

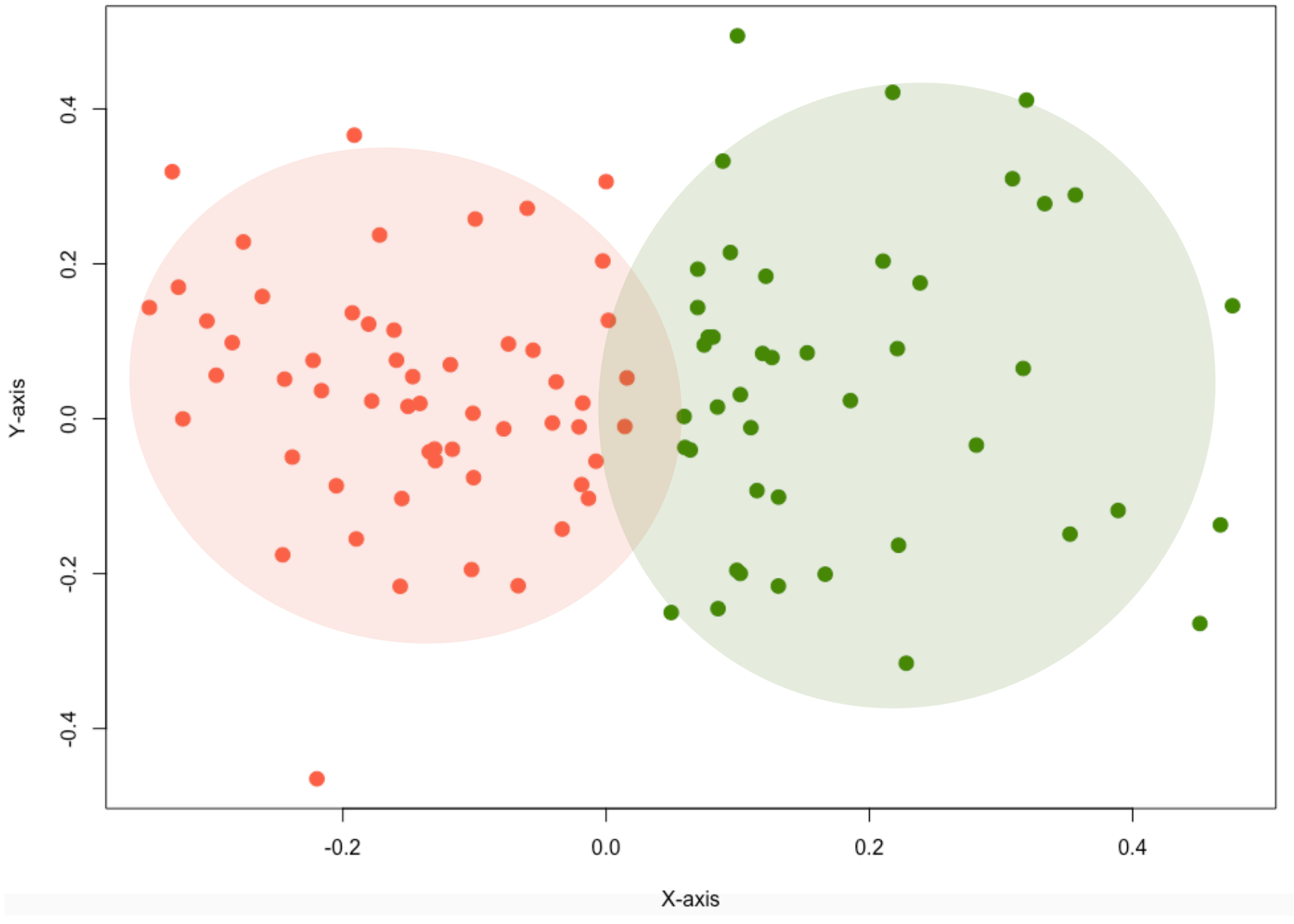
What is the best way to normalize your raw counts?

- ☐ Use relative abundance
- ☐ Total Sum Normalization [TSS]
- ☐ Rarefying
- ☐ Variance stabilizing transformation (e.g. DESeq2)
- ☐ Trimmed mean of M–values normalization (e.g. EdgeR)
- ☐ Log10
- ☐ Z
- ☐ Hellinger
- ☐ …

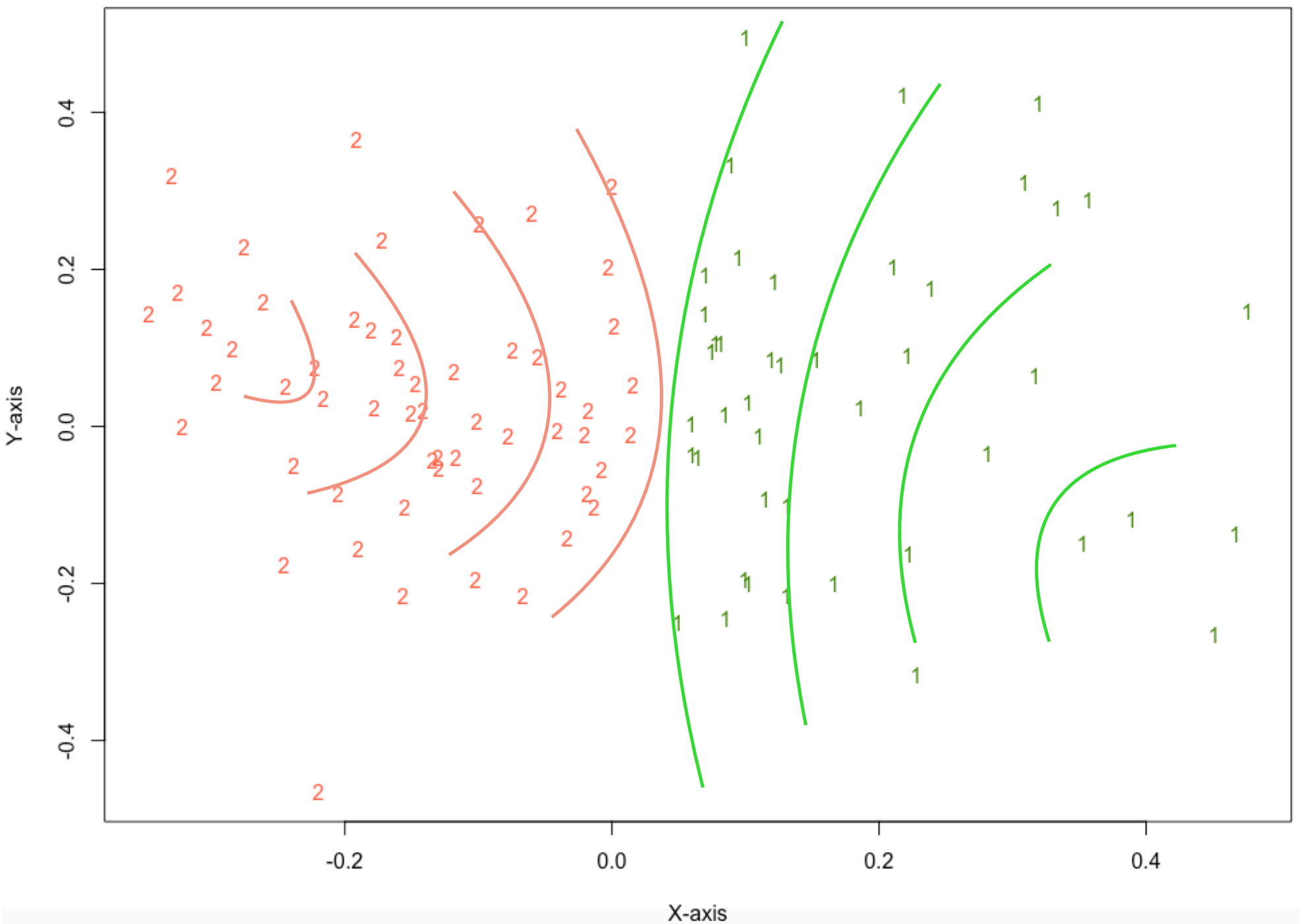
What is the best way to normalize your raw counts?

Don’t ask me? Do you have to normalize your data or would e.g. transform or rarefying your data be enough?

Do you see two clusters?



Do you see two clusters now?



PCA or PCoA?

- ☐ PCA is better.
- ☐ PCoA is better.
- ☐ Both methods can be used.
- ☐ Depense on the question.

PCA or PCoA?

You most likely work with distances and therefore you should consider a PCoA. It is possible to use a PCA but you have to standardize the abundance read data first.


MORE
THINGS
CONSIDERED