Microbiome data science with R/Bioconductor

Contents

Chapter 1

Overview

1.1 Schedule

Download the full schedule.

The schedule is summarized as follows.

- Day 1 (Tue) - Symposium; online lectures and no hands-on session
- Day 2 (Wed) - Online lectures; hands-on session on **R/Bioconductor framework**
- Day 3 (Thu) - Online lectures; hands-on session on **microbiome data analysis methods**
- Day 4 (Fri) - Online lectures; advanced microbiome **data analysis methods**

1.2 Learning goals

This course will teach the **basics of microbiome data analysis and integration with R/Bioconductor**, a popular open source environment for scientific data analysis.

You will get an overview of the reproducible data analysis workflow, with recent examples from published studies.

After the course you will know how to approach new tasks in microbiome data science by utilizing the available R tools and documentation. In particular, you understand the concepts of data containers, reproducible workflows, and standard concepts in microbiome data analysis.

1.3 Target audience

The course is primarily designed for advanced MSc and PhD students, Postdocs, and biomedical researchers who wish to learn and develop new skills in scientific programming and microbiome data science. Academic students and researchers from Finland and abroad are welcome and encouraged to apply. The course has limited capacity, and priority will given for local students.

Expected background Earlier experience with R or another programming language is expected. The teaching format allows adaptations according to the student's learning speed.

1.4 Learning material

The teaching builds on the open online tutorial, Orchestrating Microbiome Analysis (https://microbiome.github.io/OMA). The openly licensed teaching material will be available online during and after the course, following recommendations on open education.

The training material walks you through the standard steps of microbiome data analysis covering data import, processing, exploration, analysis, visualization, reproducible reporting, and best practices in open science. We teach generic data analytical skills that are applicable to common data analysis tasks encountered in modern omics research. The teaching format allows adaptations according to the student's learning speed.

Link to online Gitter chat: https://microbiome.github.io

Chapter 2

Checklist: preparing for the course

2.1 Questionnaire on the background of participants

Fill in the anonymous questionnaire.

This information will help us to understand the background of the participants better, and adjust teaching accordingly.

2.2 Installing the required R/Bioconductor packages

Install the required software in advance.

- R (it is critical to use the latest official release!)
- RStudio; choose "Rstudio Desktop" to download the latest version. For further details, check the Rstudio home page.
- Install and load the required R packages (see below)
- After a successful installation you can start with the case study examples in the training material

2.2.1 Required R/Bioconductor packages

This section shows how to install and load all required packages into the R session, if needed. Only uninstalled packages are installed.

```
# List of packages that we need from cran and bioc
cran_pkg <- c("BiocManager", "bookdown", "dplyr", "ecodist", "ggplot2",
              "gridExtra", "kableExtra",  "knitr", "scales", "vegan", "matrixStats")
bioc_pkg <- c("yulab.utils","ggtree","ANCOMBC", "ape", "DESeq2", "DirichletMultinomial", "mia", "miaViz", "mi
github_pkg <- c("miaTime")


# Get those packages that are already installed
```

```
cran_pkg_already_installed <- cran_pkg[ cran_pkg %in% installed.packages() ]
bioc_pkg_already_installed <- bioc_pkg[ bioc_pkg %in% installed.packages() ]
github_pkg_already_installed <- github_pkg[ github_pkg %in% installed.packages() ]

# Get those packages that need to be installed
cran_pkg_to_be_installed <- setdiff(cran_pkg, cran_pkg_already_installed)
bioc_pkg_to_be_installed <- setdiff(bioc_pkg, bioc_pkg_already_installed)
github_pkg_to_be_installed <- setdiff(github_pkg, github_pkg_already_installed)

# Reorders bioc packages, so that mia and miaViz are first
bioc_pkg <- c(bioc_pkg[ bioc_pkg %in% c("mia", "miaViz") ],
              bioc_pkg[ !bioc_pkg %in% c("mia", "miaViz") ] )

# Combine to one vector
packages <- c(bioc_pkg, cran_pkg)
packages_to_install <- c( bioc_pkg_to_be_installed, cran_pkg_to_be_installed, cran_pkg_to_be_installed)
```

```
# If there are packages that need to be installed, install them
if( length(packages_to_install) ) {
   BiocManager::install(packages_to_install)
}
```

Now all required packages are installed, so let's load them into the session. Some function names occur in multiple packages. That is why miaverse's packages mia and miaViz are prioritized. Packages that are loaded first have higher priority.

```
# Loading all packages into session. Returns true if package was successfully loaded.
loaded <- sapply(packages, require, character.only = TRUE)
as.data.frame(loaded)
```

2.3 Reading and support

- View the short online videos on R/Bioconductor microbiome data science tools.
- Check the Appendix chapter of the OMA book. In particular, read Chapter 15.3 on reproducible reporting.
- **You can run the workflows by simply copy-pasting the examples.** For further, advanced material, you can test and modify further examples from the book, and apply these techniques to your own data.
- Online support on installation and other matters, join us at Gitter

Chapter 3

Acknowledgments

3.1 Teachers and organizers

- Leo Lahti is the main teacher and Associate Professor in Data Science at the University of Turku, Finland, with specialization on microbiome research.
- Prof. Richa Ashma; local organizer.
- Doctoral candidate Renuka Potbhare; course assistant

3.2 Support

The course is funded by SPARC, and jointly organized by:

- Savitribai Phule Pune University, Pune, India
- Department of Computing, University of Turku, Finland
- CompLifeSci Biocity Research Program, Turku, Finland

The teaching materials have been developed with support from

- ML4microbiome COST action
- Horizon/RIA project FindingPheno
- CompLifeSci Biocity Research Program, Turku, Finland
- Turku University Foundation
- Academy of Finland

Citation We thank all developers and contributors who have contributed open resources that supported the development of the training material. Kindly cite the course material as Tuomas Borman and Leo Lahti (2022)

Bibliography

Tuomas Borman and Leo Lahti (2022). *Multi-omic data science with R/Bioconductor*.