Multi-omic data science with R/Bioconductor

Welcome to Oulu Summer School, June 2022

2022-06-17

Contents

Chapter 1

Overview

1.1 Contents and learning goals

This course will teach the **basics of biomedical data analysis with R/Bioconductor**, a popular open source environment for scientific data analysis. The participants get an overview of the reproducible data analysis workflow in modern multi-omics, with a focus on recent examples from published microbiome studies. After the course you will know how to approach new tasks in biomedical data analysis by utilizing available documentation and R tools.

The teaching will follow open online documentation created by the course teachers, extending the online book Orchestrating Microbiome Analysis (https://microbiome.github.io/OMA). The openly licensed teaching material will be available online during and after the course, following national recommendations on open education.

The training material walks you through the standard steps of biomedical data analysis covering data access, exploration, analysis, visualization, reproducible reporting, and best practices in open science. We will 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.

1.2 Schedule and organizers

The course will be organized in a live format (Flyer)

Venue University of Oulu. June 20-23, 2022.

Schedule Contact teaching daily between 9am – 5pm, including lectures, demonstrations, hands-on sessions, and breaks. A detailed schedule is available at the course website: (https://microbiome.github.io/course_2022_oulu).

Teachers and organizers

Leo Lahti is the main teacher and Associate Professor in Data Science at the University of Turku, with specialization on biomedical data analysis. Course assistants are *Tuomas Borman* (University of Turku) is one of the main developers of the open training material covered by the course and *Jenni Hekkala*, a PhD researcher at the University of Oulu, in the group of the course coordinator Docent *Justus Reunanen*.

The course is jointly organized by

- Health and Biosciences Doctoral Programme University of Oulu Graduate School
- Cancer & Translational Medicine Research Unit, University of Oulu
- Department of Computing, University of Turku, Finland
- Finnish IT Center for Science (CSC) supports the course with cloud computing services

1.3 How to apply

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 biomedical data analysis. Academic students and researchers from Finland and abroad are welcome and encouraged to apply. The course has limited capacity of max 20 participants, and priority will given for local students from Oulu.

Expected background Some earlier experience with R or another programming language is recommended. However, this can be compensated by familiarizing with the course material in advance, if necessary. The teaching format allows adaptations according to the student's learning speed.

Application

- Send a brief motivation letter to Jenni Hekkala first.last@oulu.fi
- Applications sent before May 20 will be given priority

Course fee

The course fee covers contact teaching and teaching material.

- 285 euros with registration by May 20, 2022
- 350 euros with registration after May 20, 2022
- Local students are exempted from the fee

Accommodation

Accommodation and travel costs are not included in the registration fee. For accommodation tips, see https://visitoulu.fi/en/arrival-overnight/

1.4 Acknowledgments

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)

Contact See https://microbiome.github.io

The source code of this repository is reproducible and contains the Rmd files with executable code. All files can be rendered at one go by running the file main.R. You can check the file for details on how to clone the repository and convert it into a gitbook, although this is not necessary for the training.

- Source code (github): miaverse teaching material
- Course page (html): miaverse teaching material

Chapter 2

Program

The course takes place daily from 9am – 5pm (CEST), including coffee and lunch breaks.

We expect that participants will prepare for the course in advance, see section 3. Online support is available.

The material follows open online book created by the course teachers, Orchestrating Microbiome Analysis https://microbiome.github.io/OMA. This is R/Bioconductor framework for multi-omic data science.

Figure source: Moreno-Indias *et al.* (2021) Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. Frontiers in Microbiology 12:11.

2.1 Day 1 - Open data science

Morning session

9-10 Coffee, Welcome & Practicalities

10-11 Lecture: Open & reproducible workflows

11-12 Demo & hands-on: Introduction to CSC RStudio notebook

12-13 Lunch break

Afternoon hands-on session

13-15 Demo: Data science framework

15-17 Hands-on: microbiome data summaries & exploration

17-18 Presentations & Discussion

2.2 Day 2 - Tabular data

Morning session

9-10 Lecture: Analysis & visualization of *tabular data*

10-12 Demo & hands-on: Univariate methods

12-13 Lunch break

Afternoon hands-on session

13-14 Demo: Multivariate data analysis & visualization

14-17 Hands-on: Multivariate data analysis & visualization

17-18 Presentations & Discussion

2.3 Day 3 - Multi-assay data

Morning session

9-10 Lecture: multi-omic data integration

10-12 Demo & hands-on: multi-assay data container

12-13 Lunch break

Afternoon hands-on session

13-15: Demo & hands-on: association analysis

13-17: Demo & hands-on: machine learning

17-18 Presentations & Discussion

2.4 Day 4 - Advanced topics

Morning session

9-10 Summary of the learning material

10-12 Demo & hands-on: custom data & advanced tools

12-13 Q & A session

Afternoon session

13-14 Lunch

14-16 Wrap-up

Chapter 3

Getting started

3.1 Checklist (before the course)

3.1.1 CSC Notebook

We will provide a temporary access to a cloud computing environment that readily contains the available software packages. Instructions to access the environment will be sent to the registered participants.

1. Read the instructions
2. Go to the CSC notebook frontpage
3. Login
    a. Haka login
        - If you have a Finnish university account, you should be able to login with Haka
        1. Press **Login** button from the frontpage
        2. Press **Haka** button
        3. Select right organization
        4. Enter login information
    b. CSC login
        - You can create a CSC account by following the instructions
        1. Press **Login** button from the frontpage
        2. Press **CSC** button
        3. Enter login information
    c. Special login
        - For those who cannot login with Haka or CSC account
        1. Contact Tuomas by email (first.v.last@utu.fi) if you are not able to login
        2. We give you a guest account
        3. Press **Special Login** button from the frontpage (below the **Login** button)
        4. Enter login information (username goes to **email** slot)
4. Join workspace
    a. Press **Join workspace** button (Top right corner)
    b. Enter the **Join Code** (Check your email)

5. Start session
    a. Press **ON** button
6. You can save files to **my-work** directory. They are kept stored even when the session is closed. **shared** folder is shared with all participants.

3.1.2 (Your own computer)

Setting up the system on your own computer is not required for the course but it can be useful for later use. The required software:

- R (version >4.1.0)
- RStudio; choose "Rstudio Desktop" to download the latest version. Optional but preferred. For further details, check the Rstudio home page.
- Install and load the required R packages (see Section 3.3)
- After a successful installation you can start with the case study examples in this training material

3.2 Support and resources

- We recommend to have a look at the additional reading tips and try out online material listed in Section 5.

You can run the workflows by simply copy-pasting the examples. For further, advanced material, you can test and modify further examples from the online book, and apply these techniques to your own data.

- Online support on installation and other matters, join us at Gitter

3.3 Installing and loading the required R packages

Note that the CSC/RStudio environment has readily installed setup. You may need the examples from this subsection if you are installing the environment on your own computer. If you need to add new packages, you can modify the examples below.

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")

# 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() ]
```

```
# 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)

# 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 )
```

```
# 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)
```

```
##                       loaded
## mia                     TRUE
## miaViz                  TRUE
## yulab.utils             TRUE
## ggtree                  TRUE
## ANCOMBC                 TRUE
## ape                     TRUE
## DESeq2                  TRUE
## DirichletMultinomial    TRUE
## BiocManager             TRUE
## bookdown                TRUE
## dplyr                   TRUE
## ecodist                 TRUE
## ggplot2                 TRUE
## gridExtra               TRUE
## kableExtra              TRUE
## knitr                   TRUE
## scales                  TRUE
## vegan                   TRUE
## matrixStats             TRUE
```

Chapter 4

Reproducible reporting with Rmarkdown

Reproducible reporting is the starting point for robust interactive data science. Perform the following tasks:

- If you are entirely new to Markdown, take this 10 minute tutorial to get introduced to the most important functions within Markdown. Then experiment with different options with Rmarkdown
- Create a Rmarkdown template in RStudio, and render it into a document (markdown, PDF, docx or other format). In case you are new to Rmarkdown Rstudio provides resources to learn about the use cases and the basics of Rmarkdown.
- Further examples are tips for Rmarkdown are available in the online tutorial to reproducible reporting by Dr. C Titus Brown.

Chapter 5

Study material

5.1 Online tutorial

The course will utilize material from the online book (beta version) Orchestrating Microbiome Analysis with R/Bioconductor (OMA).

We encourage to familiarize with the material and test examples already before the course.

5.2 Lecture slides

Slides (will be added).

Chapter 6

R programming, RStudio, and RMarkdown resources

- Basics of R programming: Base R
- Cheat sheets
- R graphics cookbook
- R Markdown tips

6.1 Resources for TreeSummarizedExperiment

- SingleCellExperiment
  - Publication
  - Project page
- SummarizedExperiment
  - Publication
  - Project page
- TreeSummarizedExperiment
  - Publication
  - Project page

6.2 Resources for phyloseq

- List of R tools for microbiome analysis
- phyloseq
- microbiome tutorial
- microbiomeutilities
- Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses (Callahan et al. F1000, 2016).

6.3 Further reading

- Data Analysis and Visualization in R for Ecologists by Data Carpentry
- Modern Statistics for Modern Biology. Holmes & Huber (2018) for background in statistical analysis
- Microbiome Data Science. Shetty & Lahti, 2019

Bibliography

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