Microbiome data science with R/Bioconductor

CSC course 2022

Welcome!

Target audience

- Advanced MSc, PhD & postdoctoral researchers who wish to learn new skills in scientific programming and multi-omic data analysis
- Focus on microbiome research
- Earlier experience with R is expected
- Questionnaire overview

Learning goals

- microbiome data science with R/Bioconductor, a popular open-source environment for life science informatics
- key concepts in microbiome bioinformatics
- open & reproducible data science workflow

After the course you will know how to approach new tasks in the analysis of taxonomic profiling data by taking advantage of available documentation and R tools.

Overview of the week

Day 1 Basic (microbiome) data wrangling

Day 2 Key concepts in microbiome data science

Day 3 Community-level analysis and visualization

Day 4 Advanced topics (time series, multi-omics integration)

Figure 1: Moreno-Indias et al. (2021) Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. Frontiers in Microbiology.

Daily program

09:00-09:30: Overview lecture

09:30-12:00: Practical session

- Hands-on practice with supervision
- Joint demonstration sessions

12:00-12:30: Recap, Q & A

Prerequisites

Google Doc

Questions at the end of the gdoc are welcome!

Day 1: Basic data wrangling

Time	Theme
09-10	reproducible reporting & data science workflow
10-11	data import & data containers
11-12	data wrangling basics

Time	Theme
12-	Summary, Q & A

Software & learning environment

- Temporary access to the notebook cloud environment provided by CSC with preinstalled software.
- We also encourage to test the installation on your own system; limited support for this will be available.

Acknowledgments

Lecturers:

- Leo Lahti, Assoc. Prof.
- Chouaib Benchraka, Scientific programmer

Department of Computing, University of Turku, Finland datascience.utu.fi

Organizers:

- Finnish IT Center for Science (CSC)

Turun yliopisto
University of Turku

CSC

Funding sources

Development work has received support from several sources.

Support

- Breakout rooms
- Online chat (Gitter) https://gitter.im/microbiome/miaverse
- Practical info (gdoc)
- If you need a small break, take it

Teaching material

- Teaching follows the open online book (beta version) created by the course teachers, Orchestrating Microbiome Analysis.

- The openly licensed teaching material, exercises and slides will be available online during and after the course.

Learning goals for *today*

- Set up reproducible data science workflows with Quarto

- Understand the structure of the microbiome data container

- Carry out basic data operations (e.g. subsetting, aggregation)

Getting started

- CSC notebook access OK?
- R, Rstudio, R packages installation OK?
- First task: reproducible workflow & Quarto documents (in a moment)

Questions?

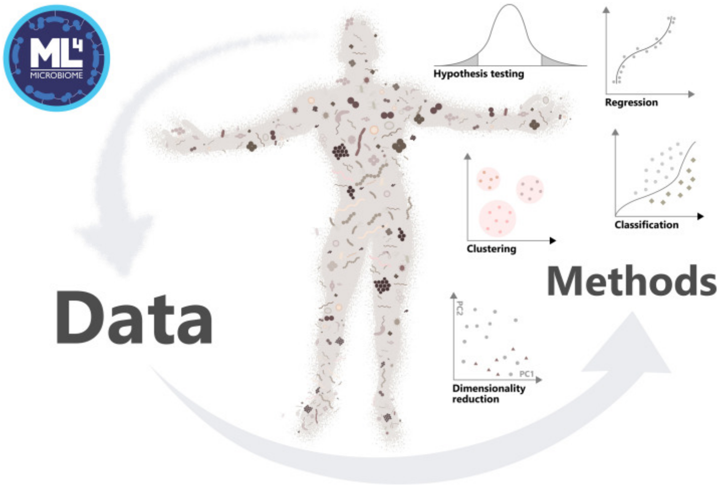
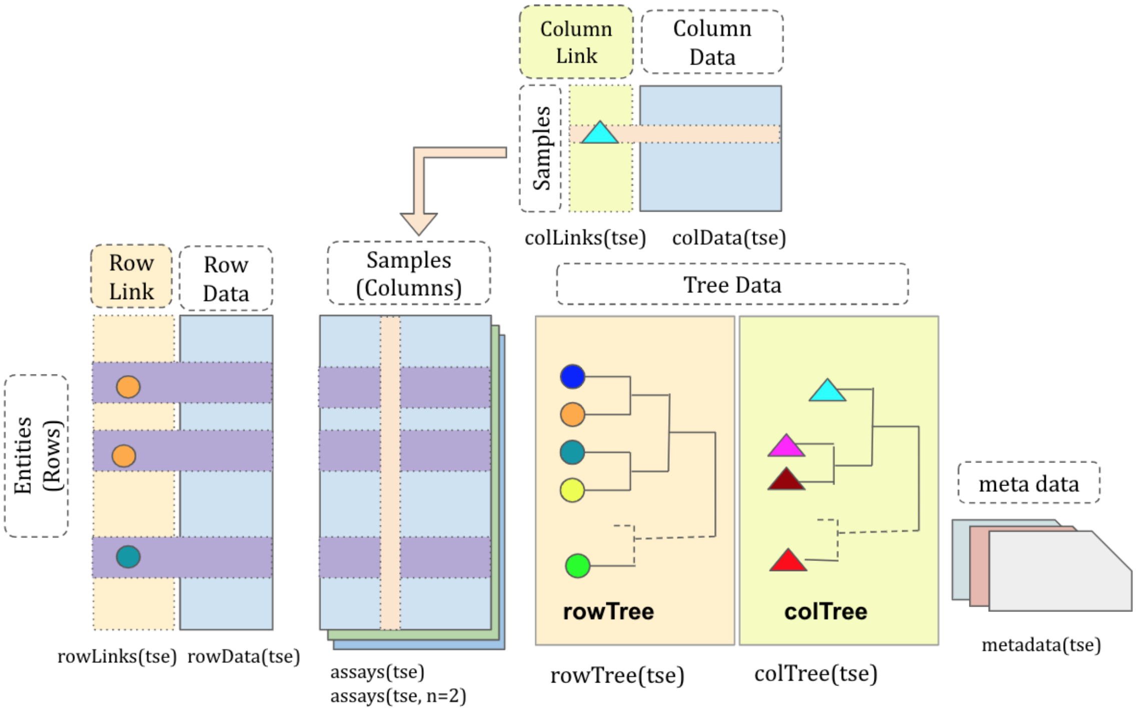


Figure 2: Moreno-Indias et al. (2021) Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. Frontiers in Microbiology.

Data containers in R/Bioconductor

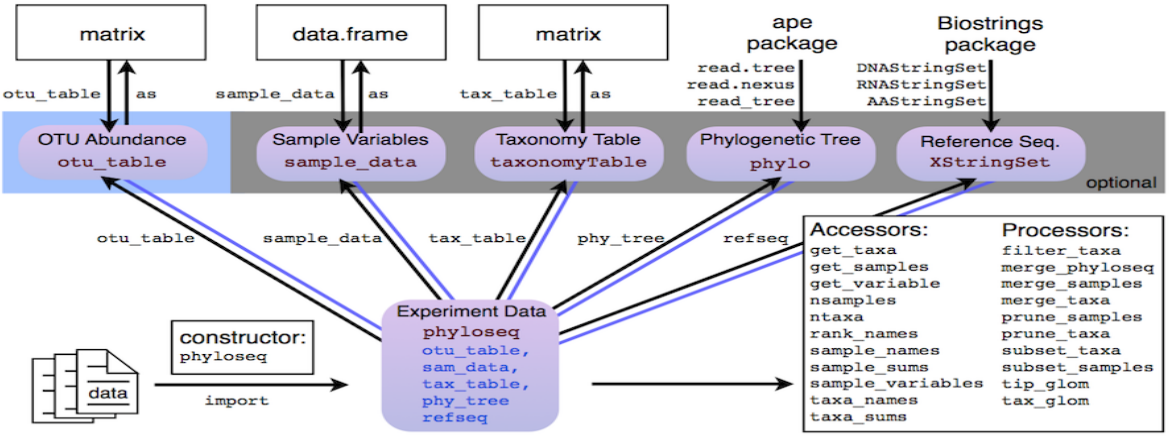
TreeSummarizedExperiment

Huang et al. F1000, 2021



Alternative data container: ***phyloseq***

Current standard for (16S) microbiome bioinformatics in R (J McMurdie, S Holmes et al.)



Demonstration data

Loading an example data set

- **Task 2: load and summarize example data (TreeSE container)**
- Troubleshooting
- Brief overview on data containers (video slides revisited)

Open microbiome data sets

- R package data (mia, miaViz, miaTime)
- Human studies: curatedMetagenomicData (Pasolli et al Nat Meth 2017)
- Other studies: microbiomeDataSets (Lahti et al.)

Task

- **Task 3: Explore TreeSE components (OMA Chapter 18.2)**
- assays, colData, rowData (trees, metadata)
- Troubleshooting
- Summary on data containers (selected video slides revisited)

Further tasks

If you complete the task fast, check out other OMA Exercises on **data containers**.

Data wrangling

Overview so far

By now, you are supposed to be able to:

- understand the basic structure of the TreeSE data container
- extract specific components from the object (assays, sample & feature info, trees)

-> How to manipulate & operate with this data object?

Basic data operations

- Subsetting

- Components

-> See the example solutions.

Transformations

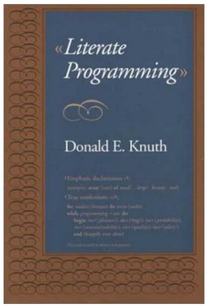
- Presence/absence
- Compositional (percentages)
- Log_{10}
- CLR and other *Aitchison* transformations
- Phylogenetic transformations (e.g. philr)
- Custom transformations

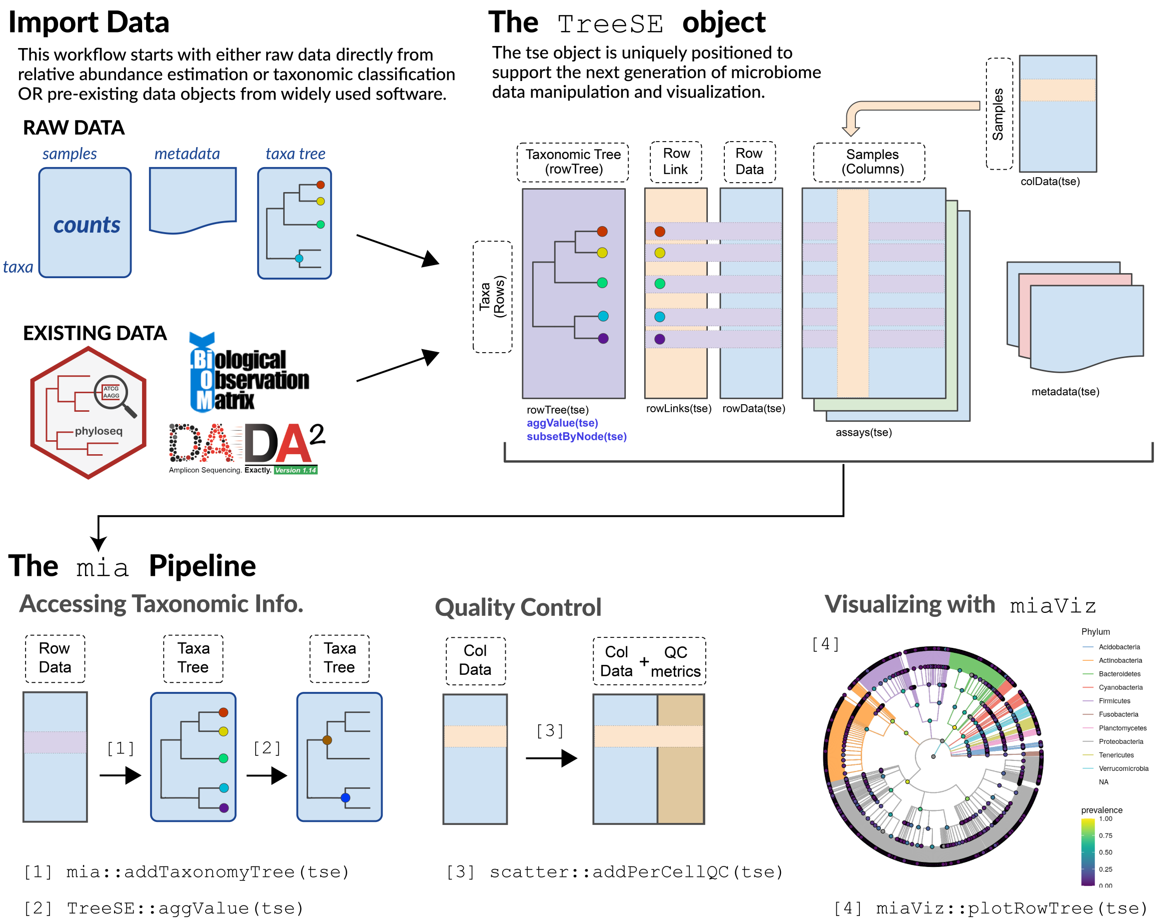
Workflow

Data containers support collaborative development of analysis methods & workflows

Understanding literate programming

Programming paradigm in which a computer program is given as an explanation of its logic in a natural language, embedded with code chunks, from which compilable source code can be generated.
(Adapted from *Wikipedia*)





This workflow starts with either raw data directly from relative abundance estimation or taxonomic classification OR pre-existing data objects from widely used software.

The tse object is uniquely positioned to support the next generation of microbiome data manipulation and visualization.

```
[1] mia::addTaxonomyTree(tse)
[2] TreeSE::aggValue(tse)
[3] scatter::addPerCellQC(tse)
[4] miaViz::plotRowTree(tse)
```

Figure 3: Domenick Braccia, EuroBioc 2020 (microbiome.github.io)