Workflow for multi-modal metabolomic and microbiome analysis in the R/Bioconductor ecosystem

Vilhelm Suksi, 41856

[vsuksi@abo.fi](mailto:vsuksi@abo.fi)

Biochemistry major, bioscience program at Åbo Akademi

Supervisor: Leo Lahti, Turun Yliopisto

2023



Abstract

Insight in how the microbiome relates to human biology is increasingly data-driven, where new data science methodologies need to account for hierarchical, heterogenous and multi-modal microbiome profiling data. The planned contribution presented herein is aimed at improving multi-modal metabolomic and microbiome analysis through integrating methods from a metabolomics R package, Notame, to the R/Bioconductor TreeSummarizedExperiment data container. The work spans implementation of the Notame methods to the TreeSummarizedExperiment container, documentation of the methods in the TreeSummarizedExperiment container context and analysis of multi-modal data to showcase the methods. This will extend the functional scope of the R/Bioconductor ecosystem to encompass methods for extracting insight from multi-modal metabolomic and microbiome data in a user-friendly, reproducible workflow. Using the workflow, substantive research efforts will be better equipped to tackle questions relating the microbiome to physiology and

Table of Contents

Abbreviations

# Introduction

With the advent of new experimental techniques, insight in life science has become increasingly reliant on sophisticated data science methodologies. This also holds true for microbiome analysis, where new data science methodologies need to account for hierarchical, heterogenous and multi-modal microbiome profiling data. The manipulation, analysis and reproducible reporting of such data is well developed in the R/Bioconductor ecosystem, focused on high-quality open research software for life science. The R/Bioconductor ecosystem can be conceptualized as consisting of data containers, R packages and a community of users and developers, who contribute to the ecosystem in an interoperable and modular fashion. The Bioconductor package repository delivers releases consisting of a set of compatible R package versions intended for compatibility only with a certain version of R, allowing for rigorous and reproducible analysis.

The notion of data containers arises from the need to organize biological data including assay matrices and meta-data such as sample descriptions and feature annotation into a single instance, facilitating the development and usage of complex analysis workflows. For example, it is possible to exclude a sample from both the meta-data and assay data in one operation, keeping the meta-data and assay data synchronized. In Bioconductor, the SummarizedExperiment family of classes provides data container solutions for various research needs. In microbiome research, the TreeSummarizedExperiment derivative of SummarizedExperiment allows for storing taxonomical information as a hierarchical tree structure. TreeSummarizedExperiment and other SummarizedExperiment derivatives also come with functions for making efficient use of the data structure for the research at hand.

Orchestration of microbiome research in the R/Bioconductor ecosystem using TreeSummarizedExperiment lineage of containers has been explored thoroughly, including basic data manipulation, transformation, exploration and quality control, taxonomic-focused tasks and machine learning. Many of the microbiome analysis tools are implemented by the mia R package, but support for in-container metabolomics data functionalities is lacking. Indeed, no Bioconductor metabolomics packages interface with the TreeSummarizedExperiment lineage of containers, except for the SDAMS package featuring a novel algorithm for differential abundance analysis. The maplet package interfaces with the TreeSummarizedExperiment lineage of containers but is not in the Bioconductor repository which limits its use in the modular and interoperable R/Bioconductor ecosystem.

The proposed work aims to rectify the above shortcomings by integrating metabolomics functionalities, chiefly from the Notame metabolomics package, with the TreeSummarizedExperiment lineage of containers. Metabolomics functionalities of interest include preprocessing, quality control and visualizations. The metabolomics interface to the TreeSummarizedExperiment lineage will expand the R/Bioconductor ecosystem to support multi-modal analysis incorporating metabolomic data. The newly integrated functionalities will be showcased in a multi-modal analysis workflow involving microbial abundance, metabolite abundance and biomarker data from the Hintikka XO dataset. The dataset was used to study whether prebiotics would reduce the negative impact of a high-fat diet in mice. The workflow is expected to benefit microbiome research efforts, ultimately knowledge of biological functions at large.

*Bakgrund/Introduktion:*Beskrivning av forskningsområdet samt det specifika området som handlas i projektet. Beskriv även de brister i information som ditt projekt syftar till att lösa.

1. Research objectives

*Målsättningar:*Frågeställningar i projektet, de specifika frågor man söker svar på i detta project

1. Research plan

*Forskningsplan:*En utförlig beskrivning av de material och metoder som används för att uppnå målsättningarna

Specify which functionalities will be included in materials & methods

Bioinformatics projects typically require a data structure with a gene expression matrix, sample descriptions and gene annotation. (perhaps copy a figure for this, from TreeSummarizedExperiment?)

Data visualization will inevitably shape interpretation and motivate the next steps of the analysis.

1. Research schedule

*Tidtabell:*Ett detaljerat schema över projektets tidtabell. Inkludera även det skriftliga avhandlingsarbetet.

1. Research synopsis

*Sammanfattning:*En kort beskrivning av de förväntade resultaten samt betydelsen av dessa.

1. References

*Litteraturförteckning:*En komplett lista över publikationer som citerats

Textdelen av forskningsplanen skall vara skriven med 1,5 radavstånd, med 12 pt skrift (Times New Roman eller motsvarande), den vänstra marginalen skall vara 4 cm bred och den högra marginalen 2,54 cm och sidorna justerade. Figurtexter kan med fördel skrivas med mindre font så som 10 pt (Times New Roman eller motsvarande). En lämplig längd på forskningsplan kan vara 10-15 sidor, men det finns inga specifika krav på längden; kom ihåg att en välskriven forskningsplan är till stor hjälp vid skrivandet av självaste pro gradu-avhandlingen.