Enhancing multi-modal metabolome and microbiome analysis with modern data containers

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 data. The planned contribution presented herein is aimed at improving multi-modal metabolome and microbiome analysis using the modern R/Bioconductor TreeSummarizedExperiment data container and the metabolomics workflow suggested by the Notame R package. The work spans implementation of the Notame workflow in the TreeSummarizedExperiment context, integration with microbial abundance data and a multi-modal analysis for contrast to earlier work. 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 pathology alike.

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 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 annotations 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 metabolomics support is underdeveloped. This is evidenced by the lack of metabolomics packages that support the TreeSummarizedExperiment lineage of containers. 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, not included in the Bioconductor repository, interfaces with the TreeSummarizedExperiment lineage of containers and provides some functionality relevant for the work at hand. Shortcomings include MS data quality control, drift correction, retention time incorporated metabolite clustering, some univariate methods and multivariate models for different downstream analyses. As such, maplet is not sufficient to implement the Notame workflow.

The proposed work aims to rectify the above shortcomings by implementing metabolomics functionalities as per the Notame package in the TreeSummarizedExperiment lineage of containers. The Notame-inspired workflow is showcased in a multi-modal metabolome and microbiome analysis. Finally, the multi-modal analysis workflow is reviewed and juxtaposed with similar data analysis approaches. 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.

# Research objectives

Can I be more specific here than in the introduction?

koodin yksinkertaisuus/lyhyys, suoritusnopeus, optioitten määrä (jotain asioita voi olla helpompi tehdä) jne.

-Reproducibility

- Reportability

- Code complexity

- Processing speed

*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.