Day 4

Leo Lahti

Overview of Day 4

Time	Theme
9-10	Differential abundance analysis
10-11	Heatmaps and other visualization techniques
11-12	Time series, multi-omic data integration
12-	Summary & Feedback

Differential abundance

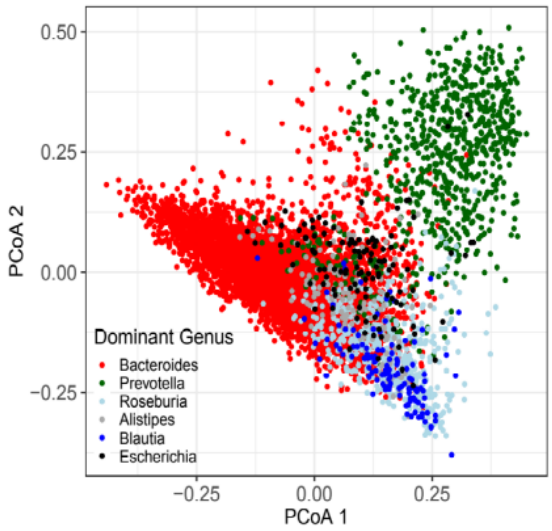
Differential abundance

- Alpha diversity: how diverse the community is?
- Beta diversity: how similar the microbial communities are?
- **Differential abundance:** how individual taxa differ between conditions?

Beta diversity revisited

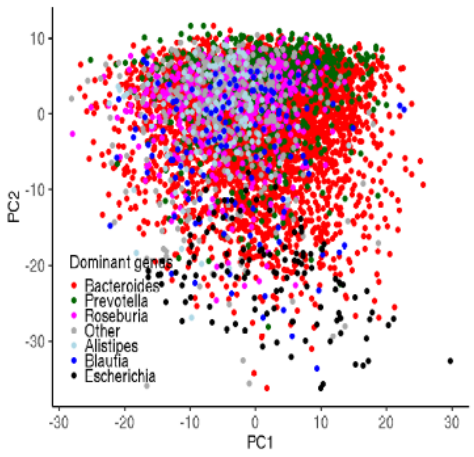
PCoA + Bray-Curtis

Preserves distances



PCA + Aitchison

Captures largest variation



Comment on this paper

Taxonomic Signatures of Long-Term Mortality Risk in Human Gut Microbiota

Aaro Salosensaari, Ville Laitinen, Aki Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, Georg Alfthan, Michael Inouye, Jeramie D. Watrous, Tao Long, Rodolfo Salido, Karenina Sanders, Caitriona Brennan, Gregory C. Humphrey, Jon G. Sanders, Mohit Jain, Pekka Jousilahti, Veikko Salomaa, Rob Knight, Leo Lahti, Teemu Niiranen

doi: https://doi.org/10.1101/2019.12.30.19015842

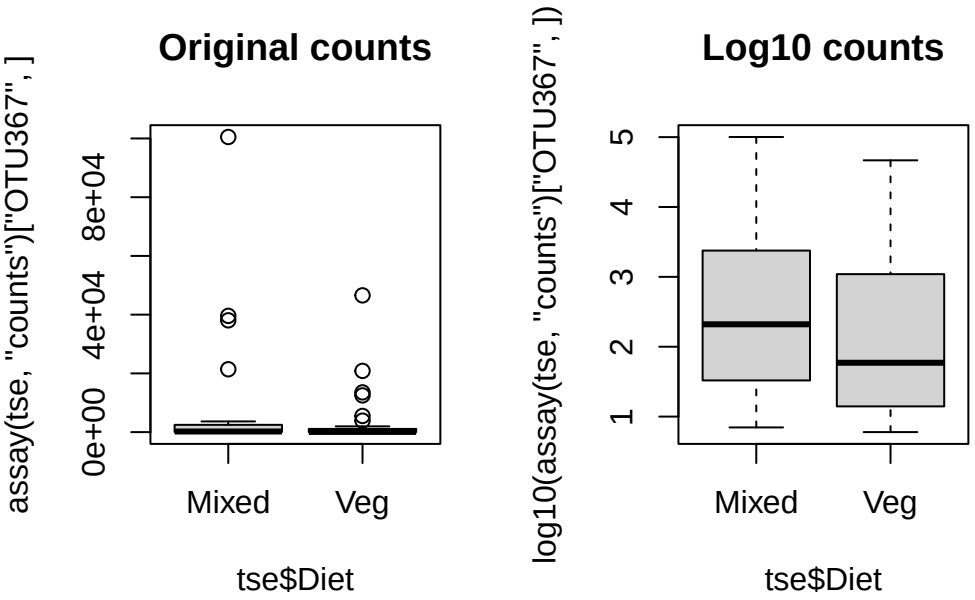
Multiple testing problem: OTU367

Also: should we transform the data for comparisons..?

```
# Load demo data
library(mia)
data(peerj13075)
tse <- peerj13075

# Compare one taxonomic unit vs. Diet
par(mfrow=c(1, 2))
boxplot(assay(tse, "counts")["OTU367",] ~ tse$Diet, main = "Original counts")
```

```
boxplot(log10(assay(tse, "counts")["OTU367",]) ~ tse$Diet, main = "Log10 counts")
```



Significance (p-value) for the row OTU367:

```
wilcox.test(assay(tse, "counts")["OTU367",] ~ tse$Diet)$p.value
```

```
Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
compute exact p-value with ties
```

```
[1] 0.1991757
```

Note: data has 674 OTUs -> multiple testing

Assumptions about the data

Taxonomic profiling data:

- Sparse
- Non-Gaussian

- Zero-inflated
- Overdispersed
- etc.

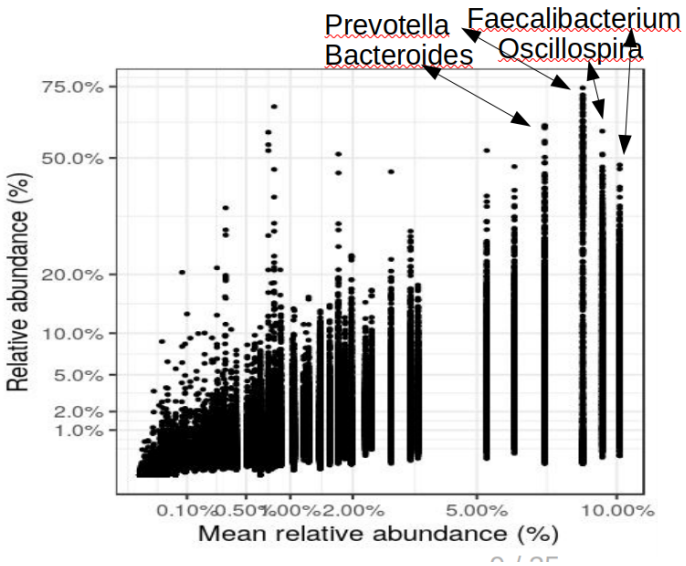
-> Usual statistical tests (Wilcoxon, t-test) can be problematic.

Taylor's law (in HITChip Atlas)

Heteroschedasticity:
Variance increases with the mean

Overdispersion:
Variance increases faster than proposed by the model

Data: HITChip Atlas



DA methods

Differential abundance (aim to) perform rigorous testing, tailored for high-throughput taxonomic profiling data with heavy multiple testing.

Some popular tools to perform DAA:

- ALDEx2
- ANCOM-BC
- LinDA
- corncob
- MaAsLin2
- DESeq2
- LEFse
- limma voom

Article | Open Access | Published: 17 January 2022

Microbiome differential abundance methods produce different results across 38 datasets

Jacob T. Nearing ✉, Gavin M. Douglas, Molly G. Hayes, Jocelyn MacDonald, Dhwani K. Desai, Nicole Allward, Casey M. A. Jones, Robyn J. Wright, Akhilesh S. Dhanani, André M. Comeau & Morgan G. I. Langille

Nature Communications **13**, Article number: 342 (2022) | Cite this article

Exercise

Test your chosen DA analysis method using the examples in OMA, Chapter 11.

- If you are uncertain, we recommend ALDEx2
- ANCOMBC supports TreeSE but is not currently installed in the notebook

ALDEx2 vs. ANCOMBC

Nearing et al. (2022):

- ALDEx2 and ANCOM-II produce the *most consistent* results across studies and agree best with the intersect of results from different approaches.
- Both ANCOM-II and ALDEx2 had the *highest precision*
- However, they suffered *lower recall values*
- The *most conservative tools*, ALDEx2 and ANCOM-II, primarily identified features that were also identified by almost all other methods.
- We recommend that researchers should use a consensus approach based on multiple DA methods to help ensure robust biological interpretations.

-> These tools are *more conservative* and have *higher precision*, with potentially *lowersensitivity*. s

ALDEx2

ALDEx2 is one well-performing DA method (see Nearing et al. 2022).

Let us first load some necessary libraries.

```
library(mia)
library(Cairo)
library(patchwork)
library(tidySummarizedExperiment)
```

```
Attaching package: 'tidySummarizedExperiment'

The following objects are masked from 'package:mia':

    full_join, inner_join, left_join, right_join

The following object is masked from 'package:XVector':

    slice

The following object is masked from 'package:IRanges':

    slice

The following object is masked from 'package:S4Vectors':

    rename

The following object is masked from 'package:matrixStats':

    count

The following object is masked from 'package:stats':

    filter
```

```
library(ALDEx2)
```

```
Loading required package: zCompositions

Loading required package: MASS


Attaching package: 'MASS'

The following object is masked from 'package:tidySummarizedExperiment':

    select

The following object is masked from 'package:patchwork':

    area

Loading required package: NADA

Loading required package: survival


Attaching package: 'NADA'

The following object is masked from 'package:IRanges':

    cor

The following object is masked from 'package:S4Vectors':

    cor

The following object is masked from 'package:stats':

    cor

Loading required package: truncnorm
```

```
library(knitr)
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --

v ggplot2 3.4.0      v purrr   0.3.5
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.4.1
v readr   2.1.3      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::bind_cols()  masks tidySummarizedExperiment::bind_cols()
x dplyr::bind_rows()  masks tidySummarizedExperiment::bind_rows()
x dplyr::collapse()   masks Biostrings::collapse(), IRanges::collapse()
x dplyr::combine()    masks Biobase::combine(), BiocGenerics::combine()
x purrr::compact()    masks XVector::compact()
x dplyr::count()      masks tidySummarizedExperiment::count(), matrixStats::count()
x dplyr::desc()       masks IRanges::desc()
x tidyr::expand()     masks S4Vectors::expand()
x dplyr::filter()     masks tidySummarizedExperiment::filter(), stats::filter()
x dplyr::first()      masks S4Vectors::first()
x dplyr::full_join()  masks tidySummarizedExperiment::full_join(), mia::full_join()
x dplyr::inner_join() masks tidySummarizedExperiment::inner_join(), mia::inner_join()
x dplyr::lag()        masks stats::lag()
x dplyr::left_join()  masks tidySummarizedExperiment::left_join(), mia::left_join()
x ggplot2::Position() masks BiocGenerics::Position(), base::Position()
x purrr::reduce()     masks GenomicRanges::reduce(), IRanges::reduce()
x dplyr::rename()     masks tidySummarizedExperiment::rename(), S4Vectors::rename()
x dplyr::right_join() masks tidySummarizedExperiment::right_join(), mia::right_join()
x dplyr::select()     masks MASS::select(), tidySummarizedExperiment::select()
x dplyr::slice()      masks tidySummarizedExperiment::slice(), XVector::slice(), IRanges::sli
```

ALDEx2

Initialize ALDEx2 analysis.

```
x <- aldex.clr(
  reads = assay(tse, "counts"),
  conds = colData(tse)$Diet,
  # 128 recommened for ttest, 1000 for rigorous effect size calculation
  mc.samples = 128,
  denom = "all",
  verbose = FALSE
)
```

ALDEx2

ALDEx2 t-test.

```
x_tt <- aldex.ttest(
  x,
  paired.test = FALSE,
  verbose = FALSE)
# effect sizes
x_effect <- aldex.effect(x, CI = TRUE, verbose = FALSE)
# combine all outputs
aldex_out <- data.frame(x_tt, x_effect)
```

ALDEx2 visualization

MA-plot

```
par(mfrow = c(1, 2))
aldex.plot(
  aldex_out,
  type = "MA",
  test = "welch",
  xlab = "Log-ratio abundance",
  ylab = "Difference",
  cutoff = 0.05
)
```

ALDEx2 summary table

```
rownames_to_column(aldex_out, "Genus") %>%
  filter(wi.eBH <= 0.05)  %>% # here we chose the wilcoxon output rather than tt
  select(Genus, we.eBH, wi.eBH, effect, overlap) %>%
  kable()
```

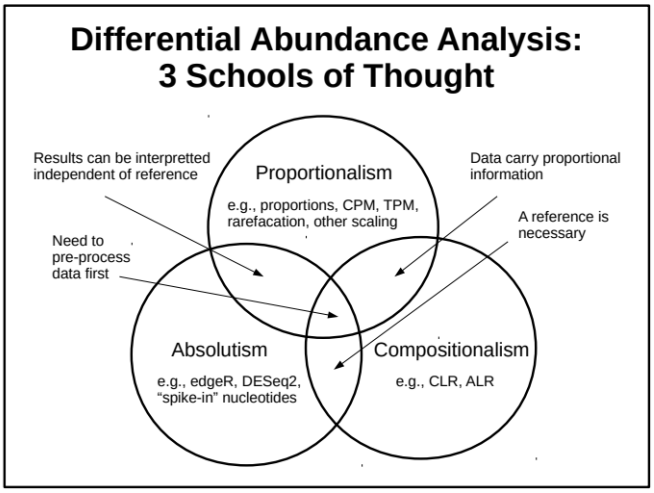


Figure 2: When it comes to differential abundance analysis, there are, generally speaking, 3 Schools of Thought. All Schools agree it is necessary to pre-process sequence count data before analysis. However, they differ in the philosophy and approach used for pre-processing. The Venn diagram identifies some beliefs held in common by the different Schools of Thought.

Figure 1: A Critique of Differential Abundance Analysis, and Advocacy for an Alternative. Thomas P. Quinn, Elliott Gordon-Rodriguez, Ionas Erb. arXiv:2104.07266 [stat.ME]

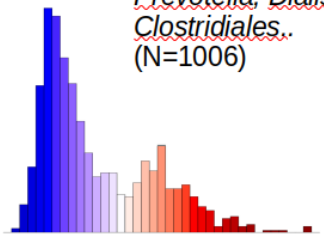
What is a suitable unit for analysis?

- Individual taxa vs. total community composition?

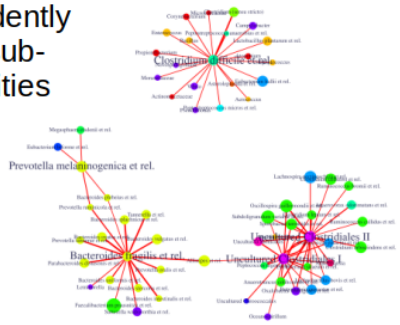
  -> Consider broader subecosystems as an intermediate between these two extremes.

Mixture models bring flexibility in modeling

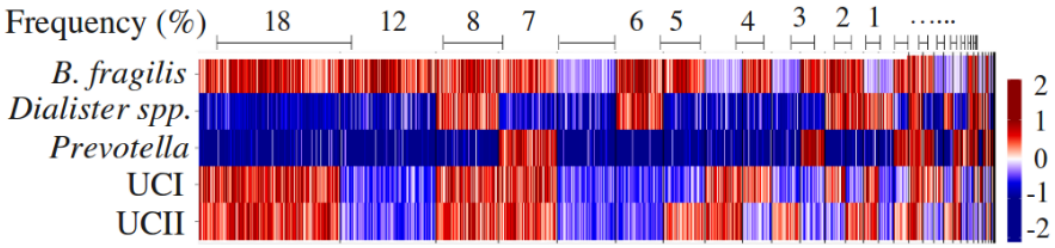
Bistable taxa:
Prevotella, *Dialister*, *Clostridiales*..
(N=1006)



Independently varying sub-communities



Ecosystem states are rich combinations of independent tipping elements ? Lahti et al. Nat. Comm. 2014



27 / 29

/ 29

A Critique of Differential Abundance Analysis, and Advocacy for an Alternative

Thomas P Quinn, Elliott Gordon-Rodriguez, Ionas Erb

It is largely taken for granted that differential abundance analysis is, by default, the best first step when analyzing genomic data. We argue that this is not necessarily the case. In this article, we identify key limitations that are intrinsic to differential abundance analysis: it is (a) dependent on unverifiable assumptions, (b) an unreliable construct, and (c) overly reductionist. We formulate an alternative framework called ratio-based biomarker analysis which does not suffer from the identified limitations. Moreover, ratio-based biomarkers are highly flexible. Beyond replacing DAA, they can also be used for many other bespoke analyses, including dimension reduction and multi-omics data integration.

Subjects: **Methodology (stat.ME)**; Genomics (q-bio.GN)
Cite as: arXiv:2104.07266 **[stat.ME]**
(or arXiv:2104.07266v2 **[stat.ME]** for this version)
https://doi.org/10.48550/arXiv.2104.07266

Summary on DA analysis

To identify individual significant taxa (w.r.t. age, diet etc.).

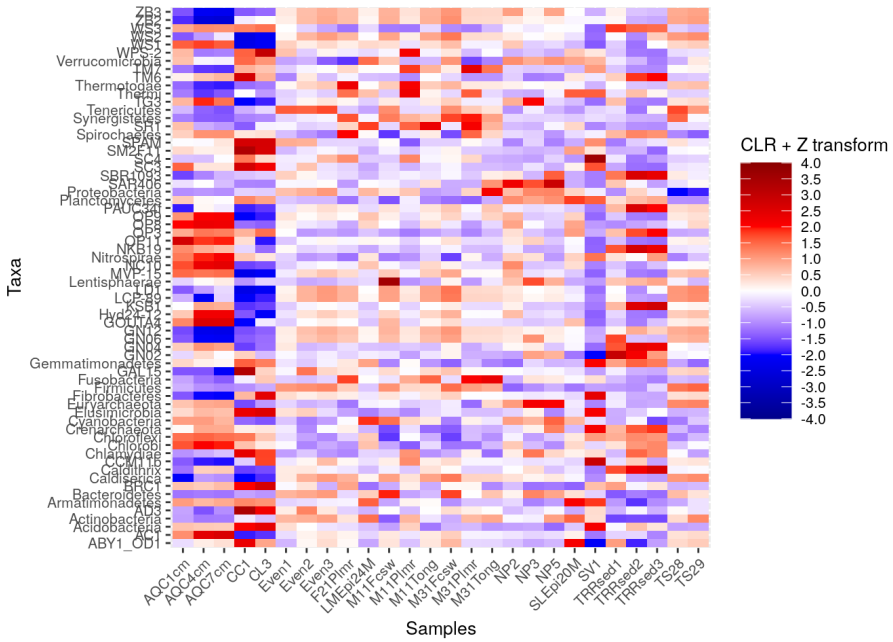
- make useful statistical assumptions for taxonomic profiling data
- control for multiple testing
- provide automated tools to summarize results (significances, effect sizes, visualizations)

-> Complements community level beta diversity analyses.

Visualization techniques

Heatmaps and other visualization techniques

Task: spend a moment testing and understanding the example visualizing taxonomic abundances on heatmaps in OMA Chapter 9



Time series

A brief demonstration on time series tools.

- miaTime
- miaSim
- miaViz

MultiAssayExperiment for multi-omic data integration

The *MultiAssayExperiment* (*MAE*) mechanism facilitates the incorporation of multiple omics (e.g. taxonomic and functional profiling).

Option	Rows (features)	Cols (samples)	Recommendation
assays	match	match	Data transformations
altExp	free	match	Alternative experiments
MultiAssay	free	free (mapping)	Multi-omic experiments

MultiAssayExperiment example data

Load example data set and list the available experiments:

```
library(mia)
data(HintikkaXOData)
mae <- HintikkaXOData
experiments(mae)
```

```
ExperimentList class object of length 3:
 [1] microbiota: TreeSummarizedExperiment with 12706 rows and 40 columns
 [2] metabolites: TreeSummarizedExperiment with 38 rows and 40 columns
 [3] biomarkers: TreeSummarizedExperiment with 39 rows and 40 columns
```

MultiAssayExperiment example data

Each element in a MultiAssayExperiment is a (Tree)SummarizedExperiment.

```
experiments(mae)[["microbiota"]]
```

```
class: TreeSummarizedExperiment
dim: 12706 40
metadata(0):
assays(1): counts
rownames(12706): GAYR01026362.62.2014 CVJT01000011.50.2173 ...
  JRJTB:03787:02429 JRJTB:03787:02478
rowData names(7): Phylum Class ... Species OTU
colnames(40): C1 C2 ... C39 C40
colData names(0):
reducedDimNames(0):
mainExpName: NULL
altExpNames(0):
rowLinks: NULL
rowTree: NULL
colLinks: NULL
colTree: NULL
```

MultiAssayExperiment example data

Each element in a MultiAssayExperiment is a (Tree)SummarizedExperiment.

```
experiments(mae)[["metabolites"]]
```

```
class: TreeSummarizedExperiment
dim: 38 40
metadata(0):
assays(1): nmr
rownames(38): Butyrate Acetate ... Malonate 1,3-dihydroxyacetone
rowData names(0):
colnames(40): C1 C2 ... C39 C40
colData names(0):
reducedDimNames(0):
mainExpName: NULL
altExpNames(0):
rowLinks: NULL
rowTree: NULL
colLinks: NULL
colTree: NULL
```

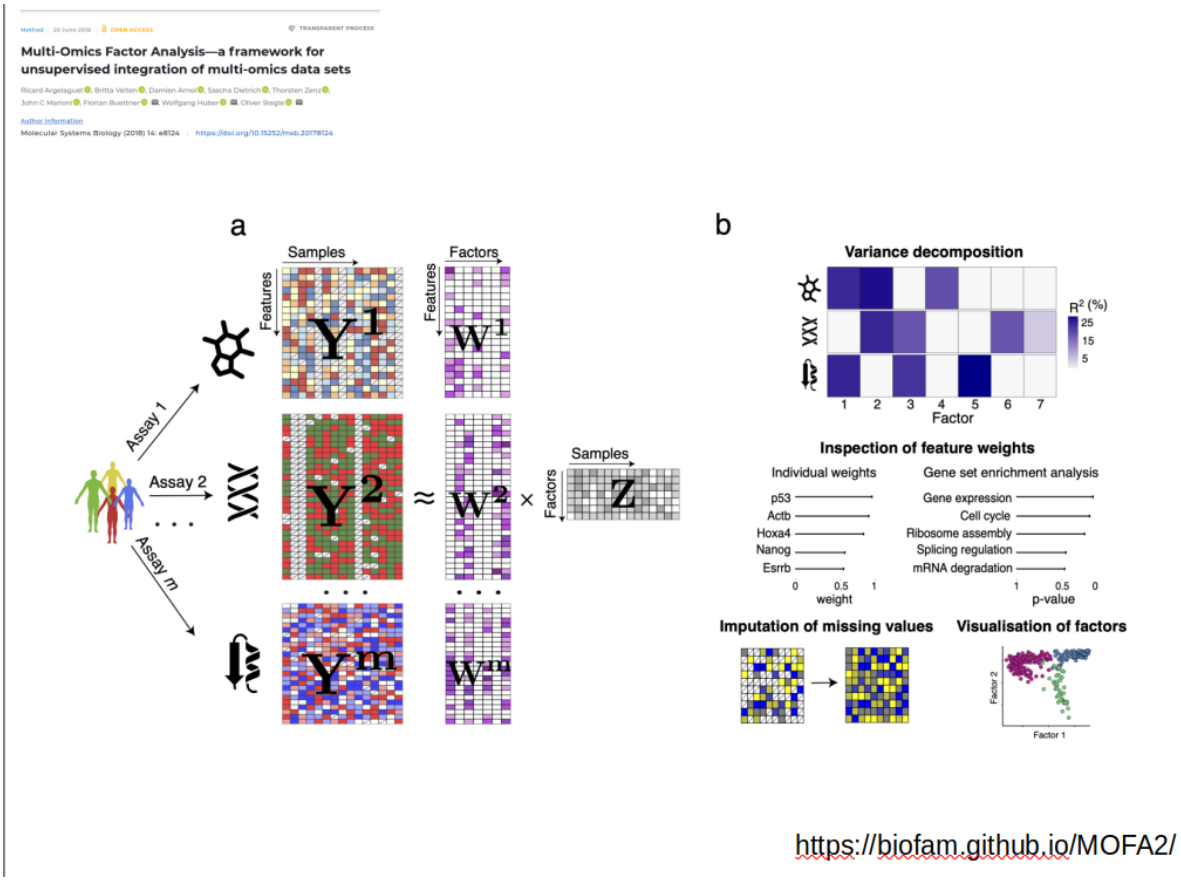
MultiAssayExperiment example data

Each element in a MultiAssayExperiment is a (Tree)SummarizedExperiment.

```
experiments(mae)[["biomarkers"]]
```

```
class: TreeSummarizedExperiment
dim: 39 40
metadata(0):
assays(1): signals
rownames(39): Triglycerides_liver CLSs_epi ... NPY_serum Glycogen_liver
rowData names(0):
colnames(40): C1 C2 ... C39 C40
colData names(0):
reducedDimNames(0):
mainExpName: NULL
altExpNames(0):
rowLinks: NULL
```

```
rowTree: NULL
colLinks: NULL
colTree: NULL
```



https://biofam.github.io/MOFA2/

REVIEW ARTICLE
Front. Genet., 28 August 2019 | https://doi.org/10.3389/fgene.2019.00627

Multitable Methods for Microbiome Data Integration

Kris Sankaran[1]* and Susan P. Holmes[2]

Property	Algorithms	Consequence
Analytical solution	Concat. PCA, CCA, ColA, MFA, PTA, Statico/Costatis	Methods with analytical solutions generally run much faster than those that require iterative updates, optimization, or Monte Carlo sampling. They tend to be restricted to more classical settings, however.
Require covariance estimate	Concat. PCA, CCA, ColA, MFA, PTA, Statico/Costatis	Methods that require estimates of covariance matrices cannot be applied to data with more variables than samples, and become unstable in high-dimensional settings.
Sparsity	SPLS, Graph-Fused Lasso, Graph-Fused Lasso	Encouraging sparsity on scores or loadings can result in more interpretable, results for high-dimensional data sets. These methods provide automatic variable selection in the multitable analysis problem.
Tuning parameters	*Sparsity*: Graph-Fused Lasso, PMD, SPLS *Number of Factors*: PCA-IV, Red. Rank Regression, Mixed-Membership CCA Prior *Parameters*: Mixed- Membership CCA, Bayesian Multitask Regression	Methods with many tuning parameters are often more expressive than those without any, since it makes it possible to adapt to different degrees of model complexity. However, in the absence of automatic tuning strategies, these methods are typically more difficult to use effectively.
Probabilistic	Mixed-Membership CCA, Bayesian Multitask Regression	Probabilistic techniques provide estimates of uncertainty, along with representations of cross-table covariation. This comes at the cost of more involved computation and difficulty in assessing convergence.
Not Normal or Nonlinear	CCpNA, Mixed-Membership CCA, Bayesian Multitask Regression	When data are not normal (and are difficult to transform to normality) or there are sources of nonlinear covariation across tables, it can be beneficial to directly model this structure.
>2 Tables	Concat. PCA, CCA, MFA, PMD	Methods that allow more than two tables are applicable in a wider range of multitable problems. Note that these are a subset of the cross-table symmetric methods.
Cross-Table Symmetry	Concat. PCA, CCA, ColA, Statico/Costatis, MFA, PMD	Cross-table symmetry refers to the idea that some methods don't need a supervised or multitask setup, where one table contains response variable and the other requires predictors. The results of these methods do not change when the two tables are swapped in the method input.

Machine learning (Random Forests)

- You can always run any ML or other analysis method available in R by just extracting individual components from TreeSE (abundance assays and sample metadata)
- For important methods, the R/Bioconductor developer community tends to come up with more integrated solutions
- The development work always benefits from feedback

Figure 5: The online version provides the text in HTML, data files and up-to-date code.

Summary

What this course was all about?

The scope of this course:

- **Brief introduction to *microbiome data science*** and some of its latest methods

This was *not* intended as a course in community ecology or machine learning in general.

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.

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

Workflow

These slides have been generated with Quarto

Overview of the week

Day 1 Basic data wrangling

Day 2 Assays, transformations & alpha diversity

Day 3 Aggregating and splitting data; beta diversity

Day 4 DA analysis, heatmaps, time series, multi-omics integration

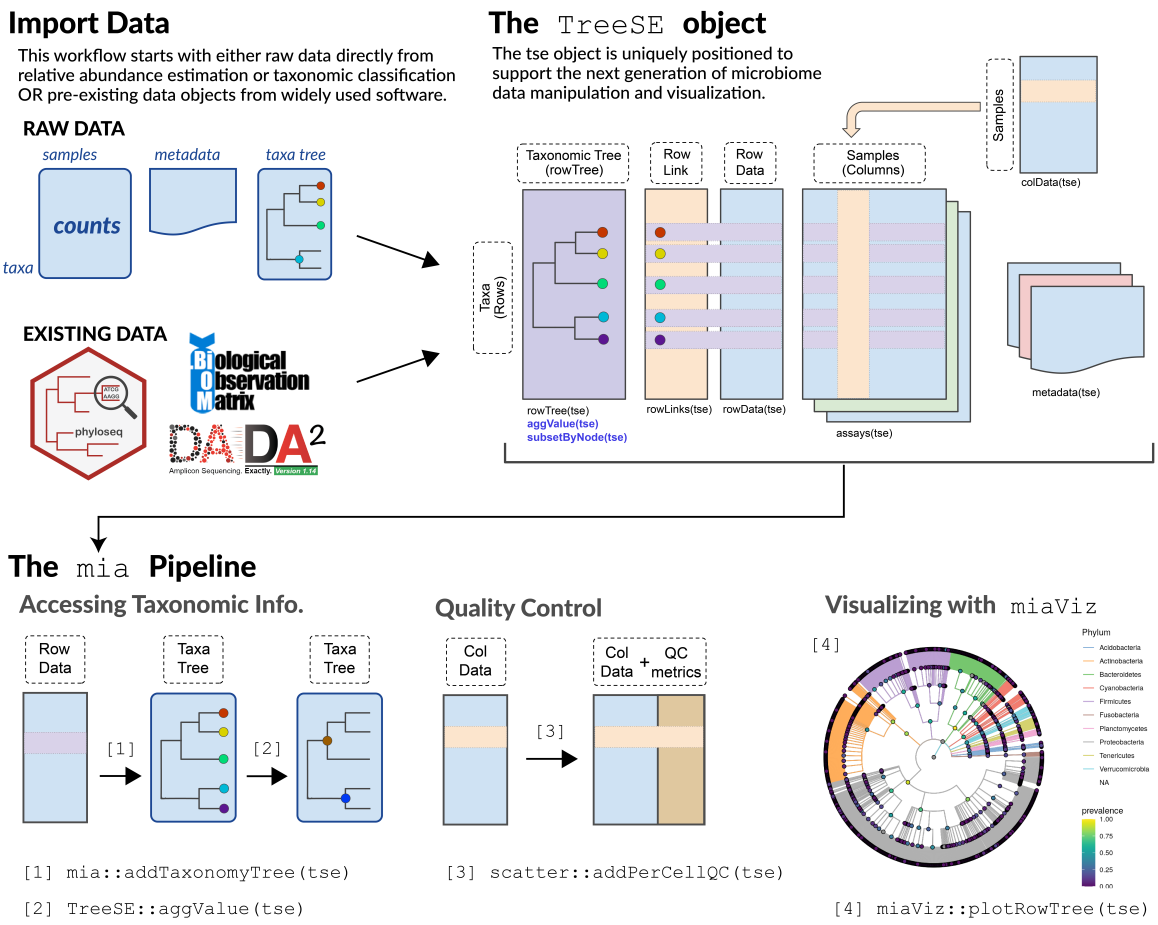
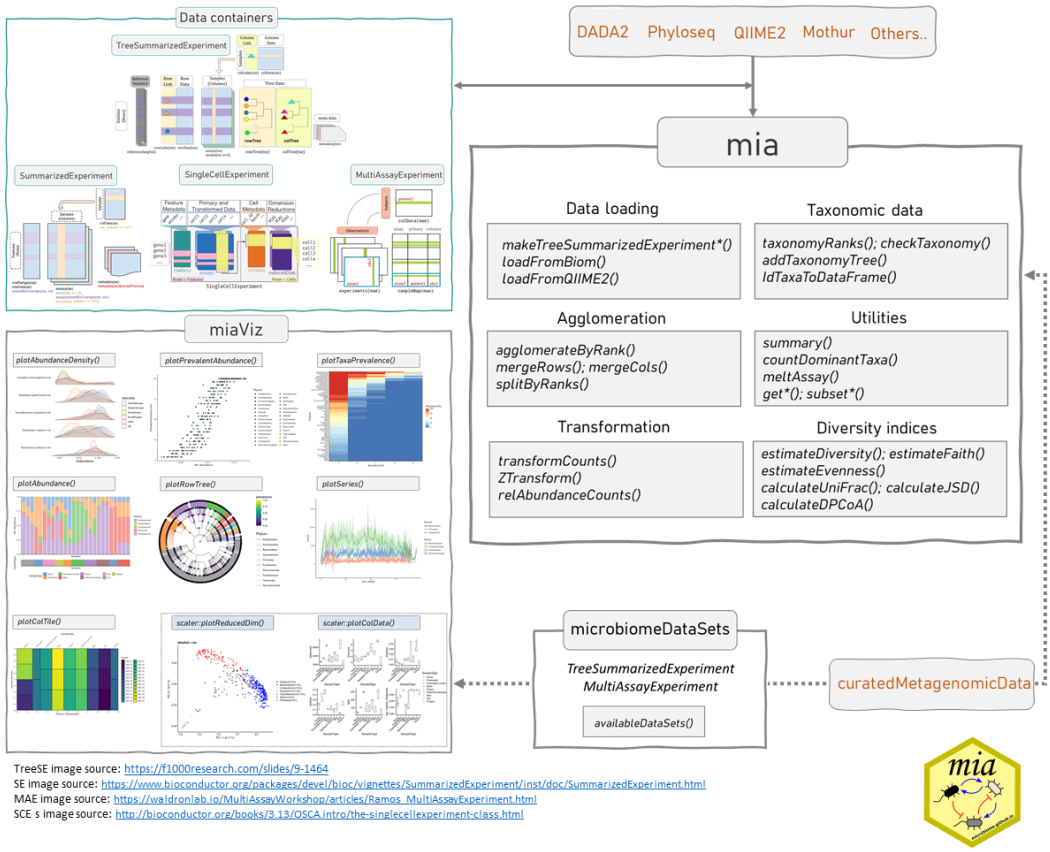


Figure 3: Domenick Braccia, EuroBioc 2020 (microbiome.github.io)



Where to find more information and examples?

OMA online book (beta version!)

Function references and vignettes in individual packages:

- mia: basic manipulation
- miaViz: visualization
- miaTime: time series
- miaSim: simulation
- philr: phylogenetic balances (Silverman et al)
- ANCOMBC (Lin & Peddada)
- curatedMetagenomicData; (Pasolli et al.) human demo data
- microbiomeDataSets; other demo data
- SingleCell tool family

- SE tool family
- ...

Join the community

Support the development and documentation by actively providing feedback & suggestions.

- Online support chat (Gitter) https://gitter.im/microbiome/miaverse
- Bioconductor Slack (#miaverse channel)
- Mailing list (see microbiome.github.io)
- Github issues (e.g. OMA issues)

Thanks!

- All participants!
- Organizer: Finnish IT Center for Science (CSC)
- Developers:
  - Leo Lahti, Assoc. Prof.
  - Tuomas Borman, PhD researcher
  - Chouaib Benchraka, Scientific programmer
  - For a full list of key developers and contributors, see OMA acknowledgments

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

Feedback

- CSC will send a feedback form

Q & A session