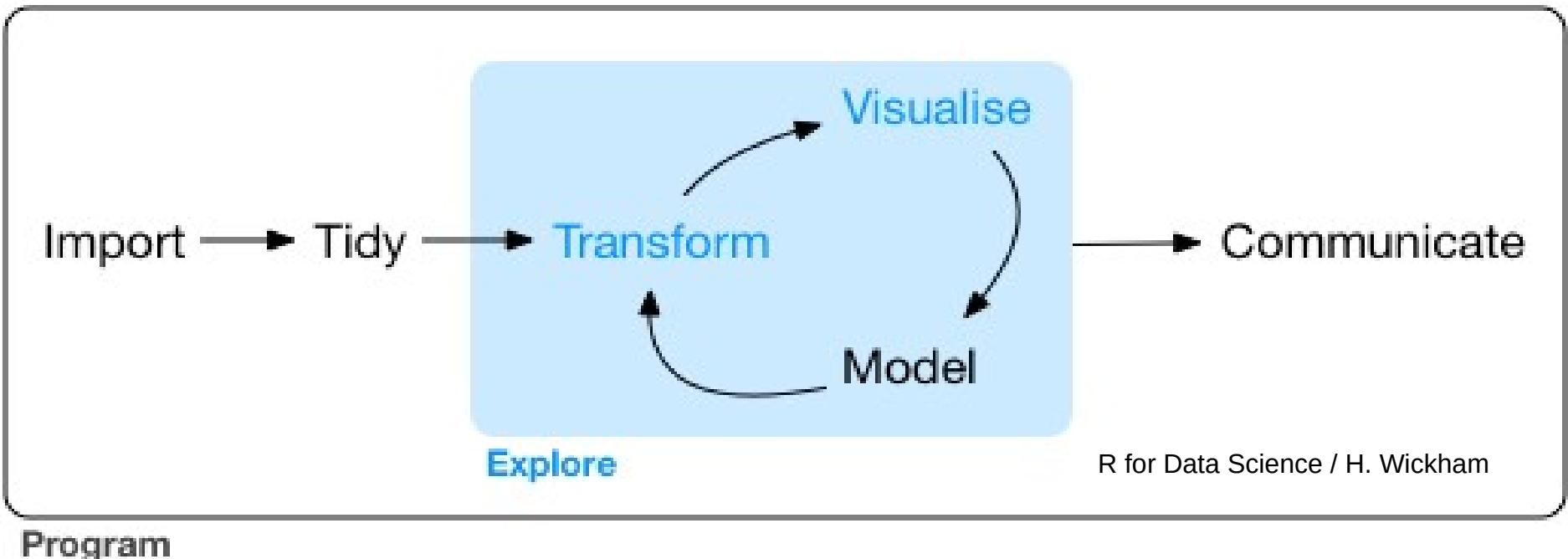




# Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

# Data science workflow



**REVISED** Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses [version 2; peer review: 3 approved]

Ben J. Callahan<sup>1</sup>, Kris Sankaran<sup>1</sup>, Julia A. Fukuyama<sup>1</sup>, Paul J. McMurdie<sup>2</sup>, Susan P. Holmes



This article is included in the [Bioconductor](#) gateway.

# microbiome R package

chat on gitter

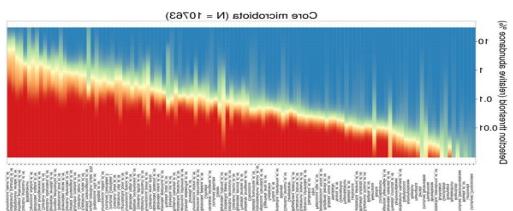
build passing

codecov 24%

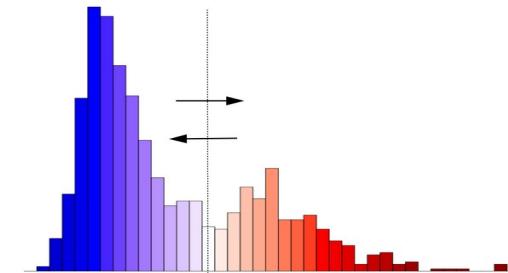
PRs welcome

## Core & prevalence

prevalence(x)  
core(x)  
core\_members(x)



## Stability & resilience



## Transformations

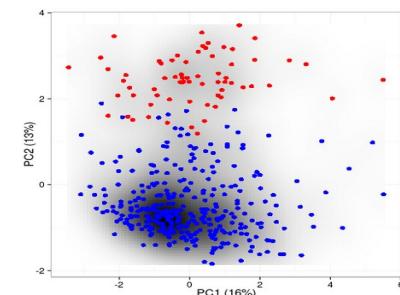
transform(x, "compositional")  
transform(x, "clr")  
transform(x, "log10p")  
transform(x, "hellinger")  
transform(x, "identity")

## Community

- Online tutorials
- Mailing list
- Gitter chat
- Example data
- Workshops

## Alpha & beta diversity

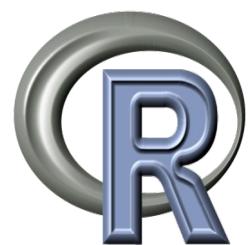
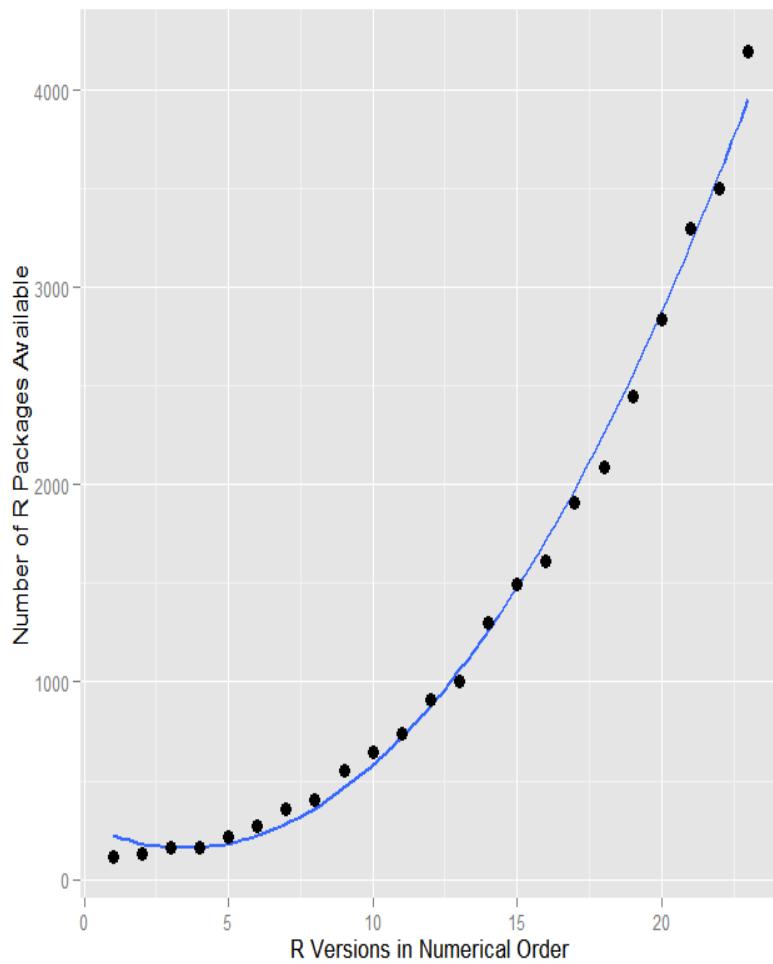
alpha(x)  
diversity(x)  
evenness(x)  
dominance(x)  
rarity(x)  
readcount(x)



## Quality control

- continuous integration
- unit tests

# Number of open analysis tools has grown exponentially



**R**OpenSci

1. Ampvis2 Tools for visualising amplicon sequencing data
2. CCREPE Compositionality Corrected by PErmutation and REnormalization
3. DADA2 Divisive Amplicon Denoising Algorithm
4. DESeq2 Differential expression analysis for sequence count data
5. edgeR empirical analysis of DGE in R
6. mare Microbiota Analysis in R Easily
7. Metacoder An R package for visualization and manipulation of community taxonomic diversity data
8. metagenomeSeq Differential abundance analysis for microbial marker-gene surveys
9. microbiome R package Tools for microbiome analysis in R
10. MINT Multivariate INTegrative method
11. mixDIABLO Data Integration Analysis for Biomarker discovery using Latent variable approaches for 'Omics studies
12. mixMC Multivariate Statistical Framework to Gain Insight into Microbial Communities
13. MMint Methodology for the large-scale assessment of microbial metabolic interactions (MMint) from 16S rDNA data
14. pathostat Statistical Microbiome Analysis on metagenomics results from sequencing data samples
15. phylofactor Phylogenetic factorization of compositional data
16. phylogeo Geographic analysis and visualization of microbiome data
17. Phyloseq Import, share, and analyze microbiome census data using R
18. qilmer R tools compliment qlime
19. RAM R for Amplicon-Sequencing-Based Microbial-Ecology
20. ShinyPhyloseq Web-tool with user interface for Phyloseq
21. SigTree Identify and Visualize Significantly Responsive Branches in a Phylogenetic Tree
22. SPIEC-EASI Sparse and Compositionally Robust Inference of Microbial Ecological Networks
23. structSSI Simultaneous and Selective Inference for Grouped or Hierarchically Structured Data
24. Tax4Fun Predicting functional profiles from metagenomic 16S rRNA gene data
25. taxize Taxonomic Information from Around the Web
26. labdsv Ordination and Multivariate Analysis for Ecology
27. Vegan R package for community ecologists
28. igraph Network Analysis and Visualization in R
29. MicrobiomeHD A standardized database of human gut microbiome studies in health and disease *Case-Control*
30. Rhea A pipeline with modular R scripts
31. microbiomeutilities Extending and supporting package based on microbiome and phyloseq R package
32. breakaway Species Richness Estimation and Modeling

# A survey for 16S

[Github.com/microsud/  
Tools-Microbiome-Analysis](https://github.com/microsud/Tools-Microbiome-Analysis)



[Journal of Biosciences](#)

October 2019, 44:115 | [Cite as](#)

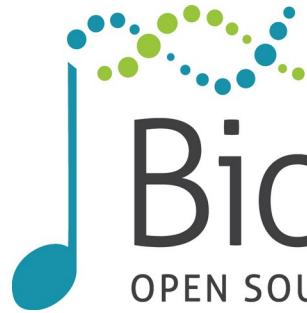


## Microbiome data science

Authors

Authors and affiliations

Sudarshan A Shetty, Leo Lahti



# Bioconductor

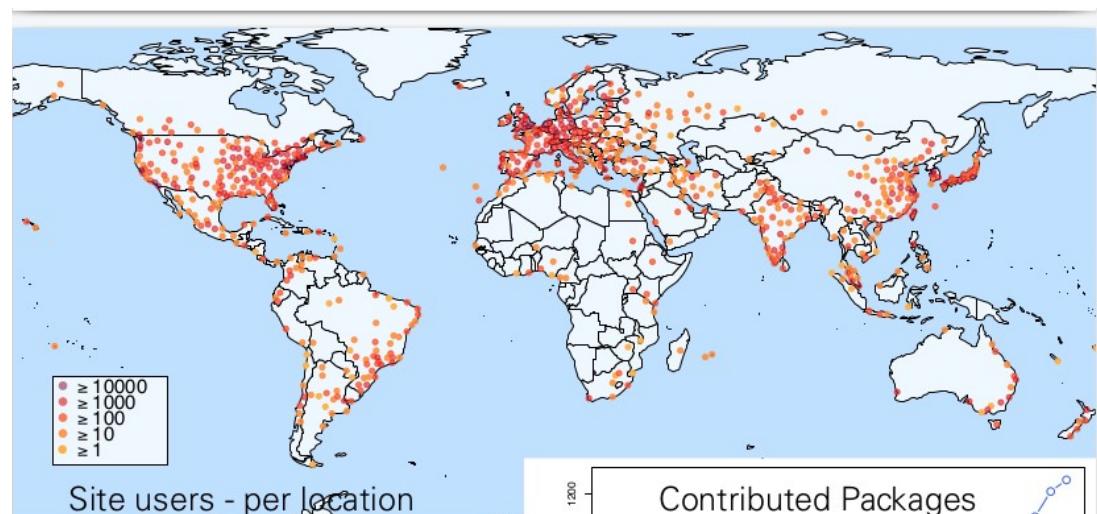
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS



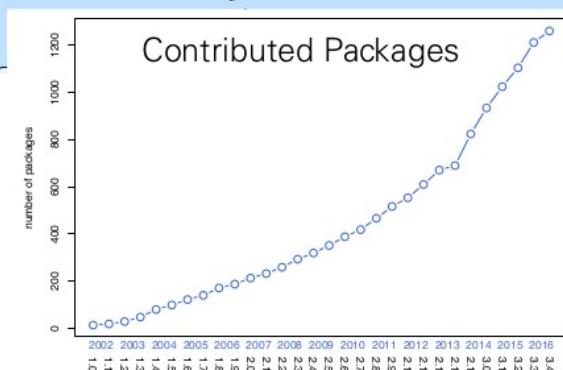
Started in 2001 as a platform for microarray data analysis

Now 2140+ packages:

- Sequencing (DNA, RNA, ChIP, SC)
- Microarrays
- Flow cytometry
- Proteomics
- Multi-omics data integration



World largest bioinformatics project  
10,000s users  
>18,000 papers in PubmedCentral



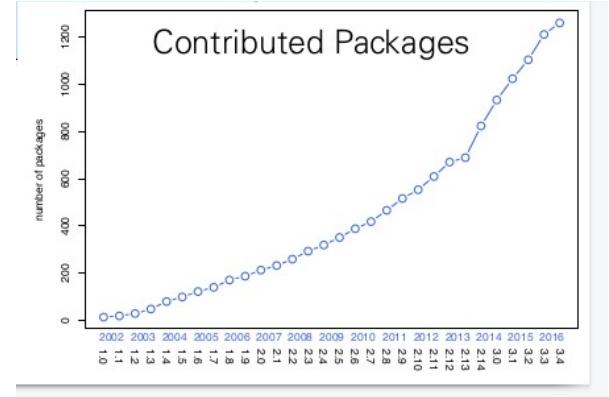
What is



?

## Collaborative research software development

- software & data repository
- tutorials & documentation
- bioinformatics peer support



50,000+ papers in PubMed



## Management:

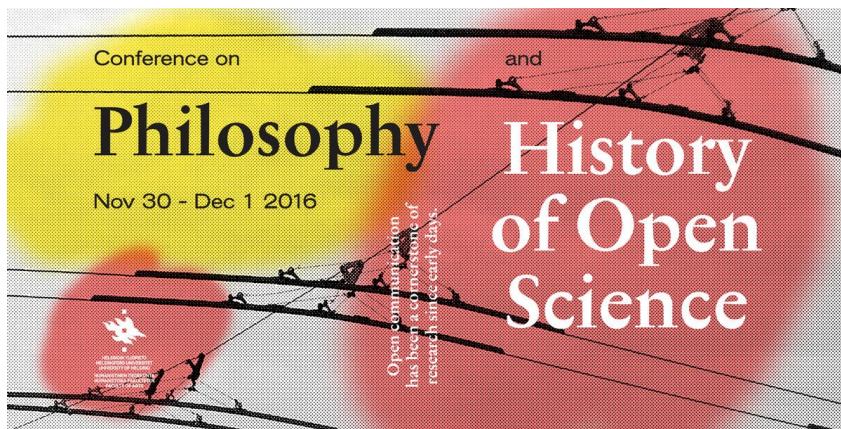
- Core team
- SAB & TAB & CAB
- Global contributor base

What marks out modern science is not the conduct of experiments (alchemists conducted plenty of experiments), but the formation of a *critical community capable of assessing discoveries and replicating results*.

The Invention of Science: A New History of the Scientific Revolution, by David Wootton

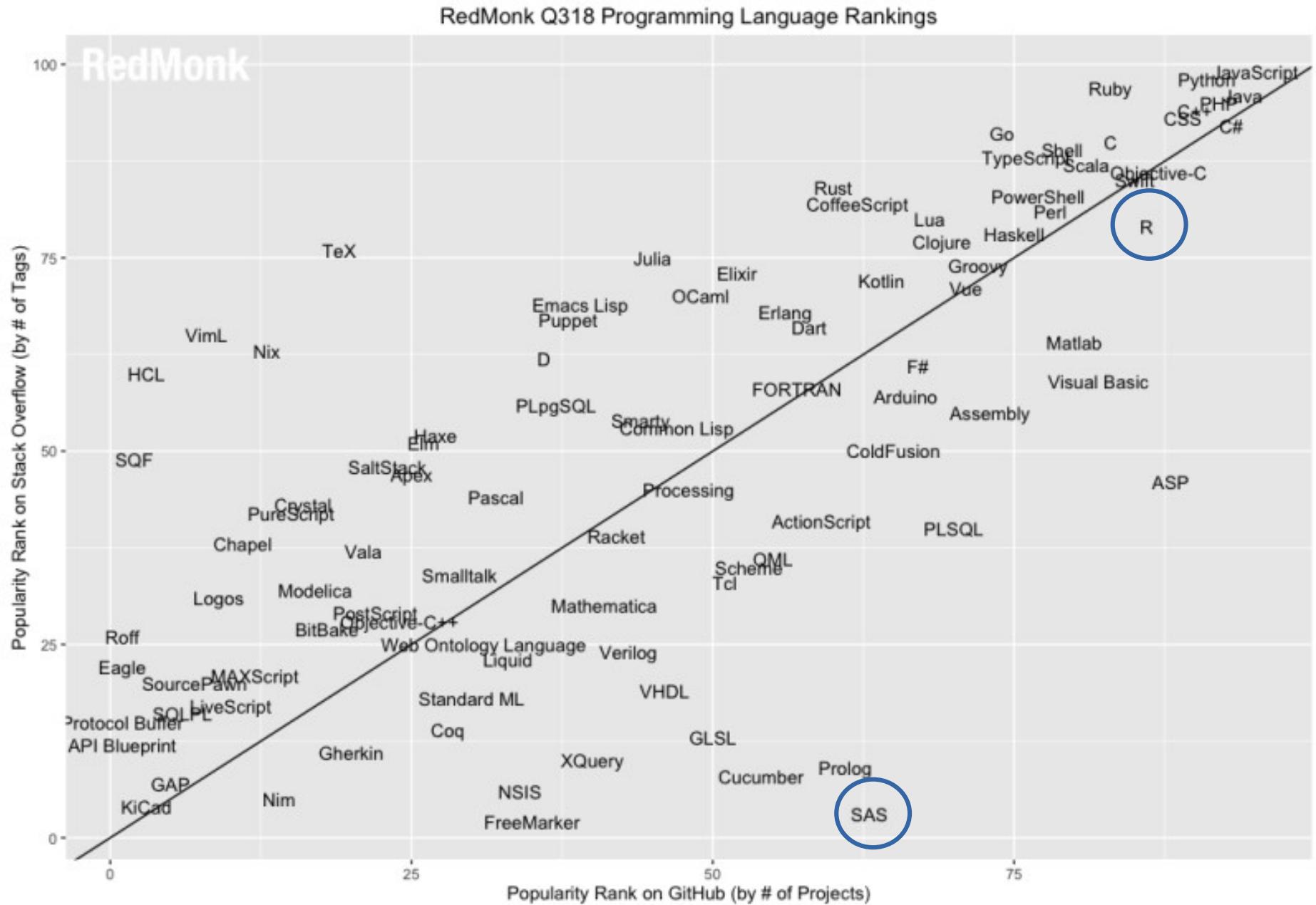
Alchemy & algorithms: perspectives on the philosophy and history of open science

▼ Leo Lahti, Filipe da Silva, Markus Petteri Laine, Viivi Lähteenaja, Mikko Tolonen



A family of alchemists at work, an engraving by Philip Galle, after a painting by Pieter Bruegel the Elder, published by Hieronymus Cock, c.1558.

# Varying cultures of open collaboration



# open data science ecosystems

**mothur**

Download Wiki Forum Blog Rifications facebook

Welcome to the website for the mothur project, initiated by Dr. Patrick Schloss and his research group at the Department of Microbiology & Immunology at the University of Michigan. This project seeks to develop a single piece of open-source, expandable software to fill the bioinformatics needs of the microbial ecology community. mothur is a command-line version of mothur, which had accelerated versions of the popular DOTUR and SONS programs. mothur has gone on to become one of the most cited bioinformatics tool for analyzing 16S rRNA gene sequences. Step inside the wiki and user forum and learn how you can use mothur to process data generated by Sanger, Pacific, Ion, 454, and Illumina (MiSeq). If you would like to contribute code to the project feel free to download the source code and make your own improvements. Alternatively, if you have an idea or a need, but lack the programming expertise, let us know through the forum and we'll add it to the queue of features we would like to add.

Subscribe to the mothur mailing list

mailing list  
Subscribe

Department of Microbiology & Immunology  
The University of Michigan Medical School  
The University of Michigan

This site is maintained by Pat Schloss  
© 2008-2019

**qiime2**

QIIME 2™ is a next-generation microbiome bioinformatics platform that is extensible, free, open source, and community developed.

Code of Conduct » Citing QIIME 2 » Learn more »

Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!

Interactively explore your data with beautiful visualizations that provide new perspectives.

Easily share results with your team, even those members without QIIME 2 installed.

Plugin-based system — your favorite microbiome methods all in one place.



PeerJ >

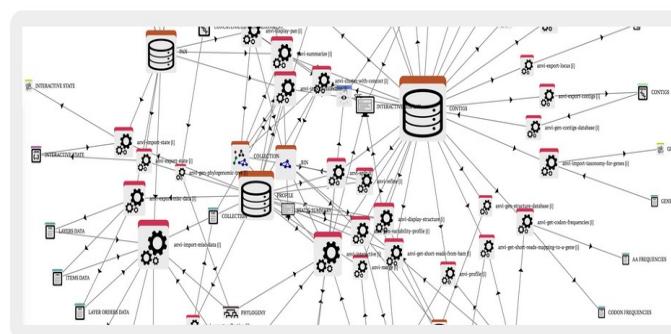
## Anvi'o: an advanced analysis and visualization platform for 'omics data

Research article Bioinformatics Biotechnology Computational Biology Genomics Microbiology

A. Murat Eren<sup>✉ 1,2</sup>, Özcan C. Esen<sup>1</sup>, Christopher Quince<sup>3</sup>, Joseph H. Vineis<sup>1</sup>, Hilary G. Morrison<sup>1</sup>, Mitchell L. Sogin<sup>1</sup>, Tom O. Delmont<sup>1</sup>

Published October 8, 2015

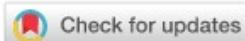
### Anvi'o in a nutshell



Anvi'o is an open-source, community-driven analysis and visualization platform for 'omics data.

# Software for the Integration of Multiomics Experiments in Bioconductor FREE

Marcel Ramos; Lucas Schiffer; Angela Re; Rimsha Azhar; Azfar Basunia; Carmen Rodriguez; Tiffany Chan; Phil Chapman; Sean R. Davis; David Gomez-Cabrero; Aedin C. Culhane; Benjamin Haibe-Kains; Kasper D. Hansen; Hanish Kodali; Marie S. Louis; Arvind S. Mer; Markus Riester; Martin Morgan; Vince Carey; Levi Waldron [✉](#)



+ Author & Article Information

Cancer Res (2017) 77 (21): e39–e42.

<https://doi.org/10.1158/0008-5472.CAN-17-0344> Article history

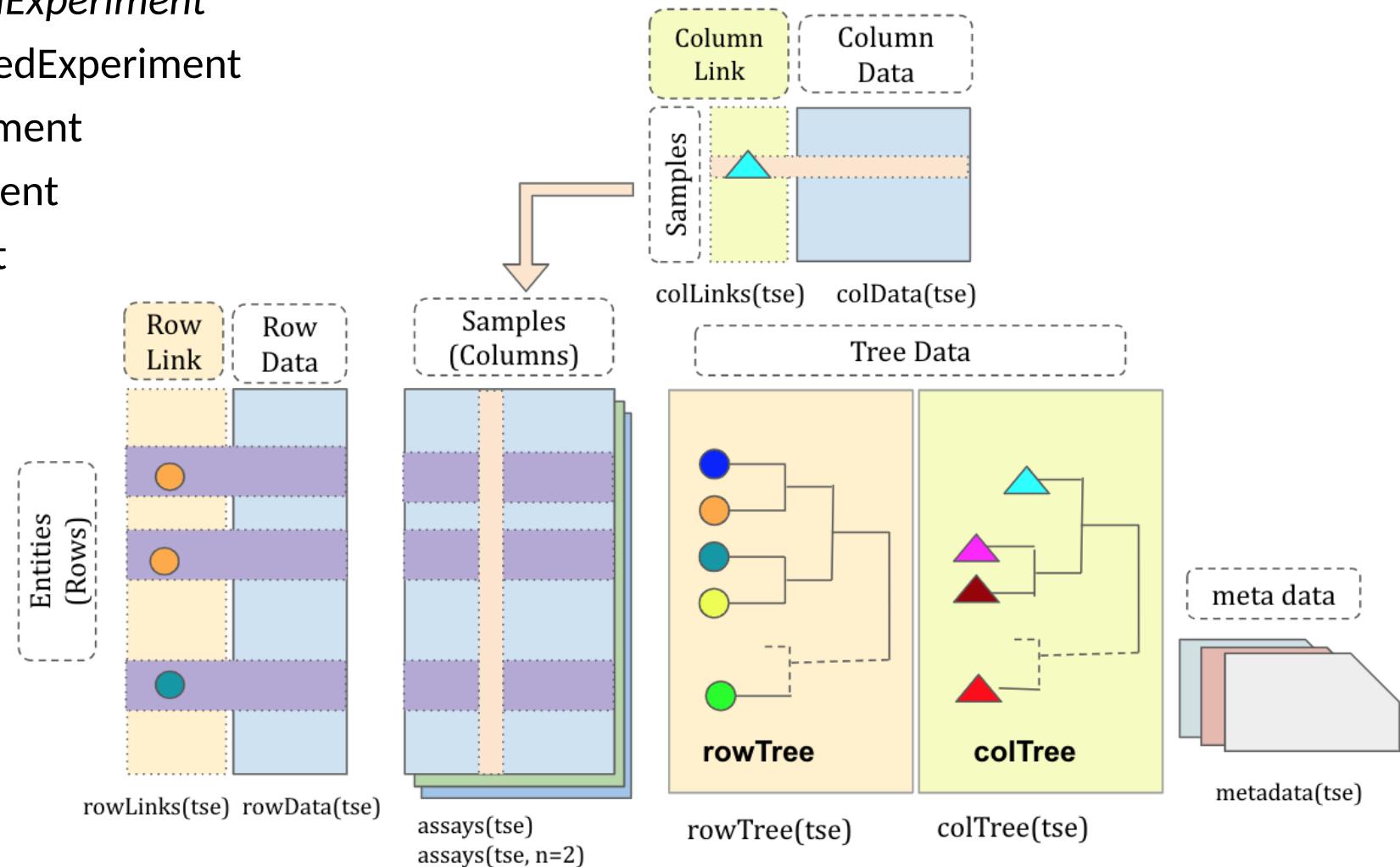
## (Tree)SummarizedExperiment

### RangedSummarizedExperiment

### MultiAssayExperiment

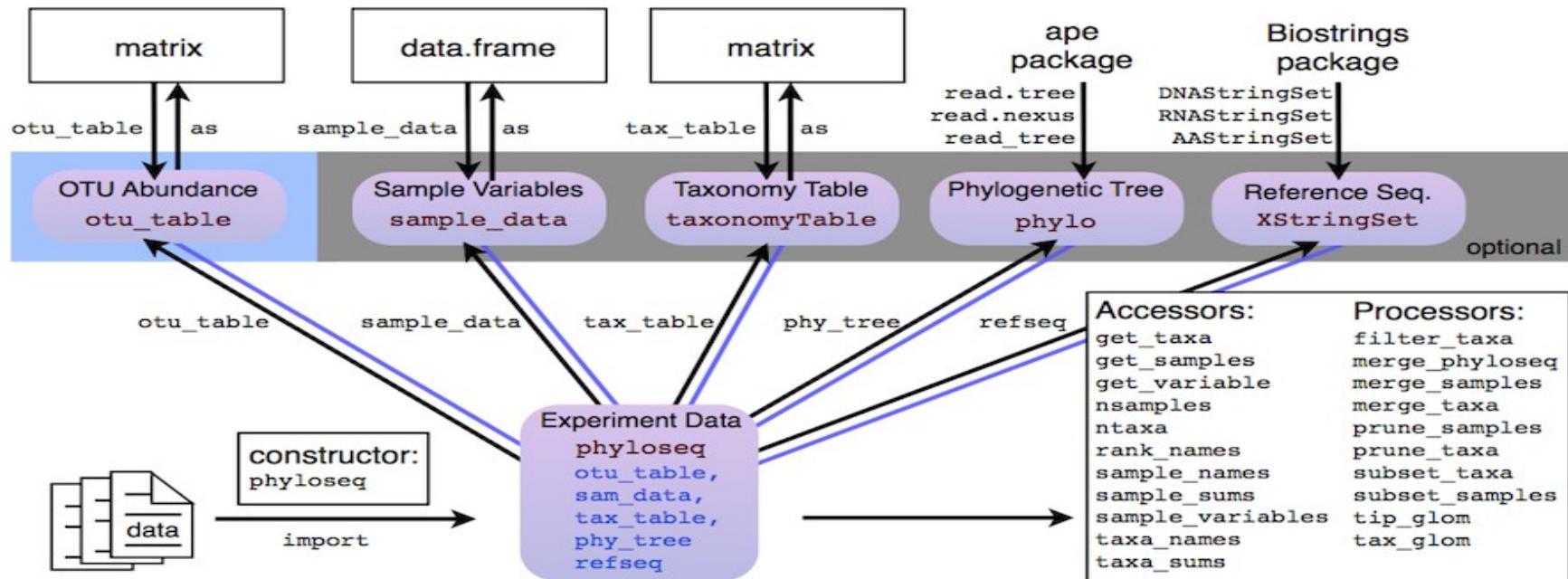
### SingleCellExperiment

### SpatialExperiment



# Alternative data container: *phyloseq*

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

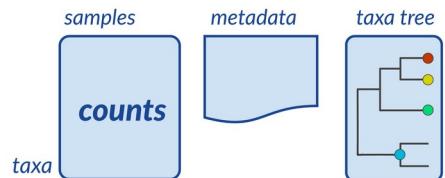


# Data containers support collaborative development of analysis methods & workflows

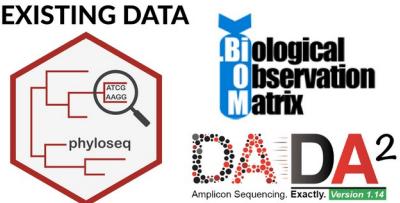
## Import Data

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

### RAW DATA

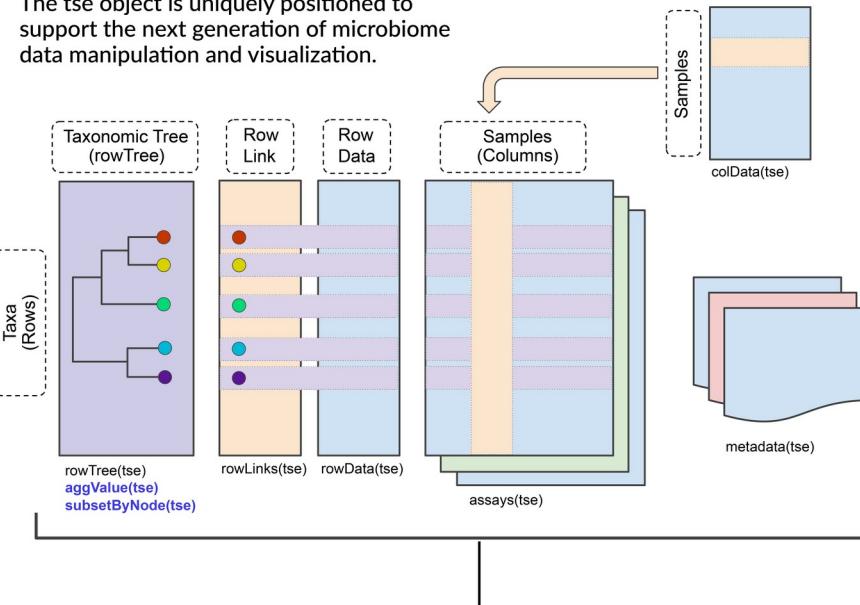


### EXISTING DATA



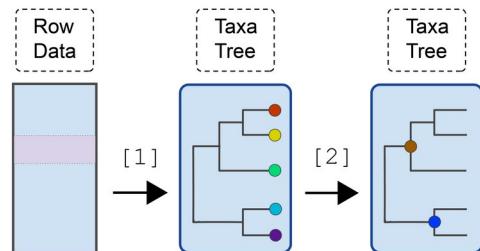
## The TreeSE object

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



## The mia Pipeline

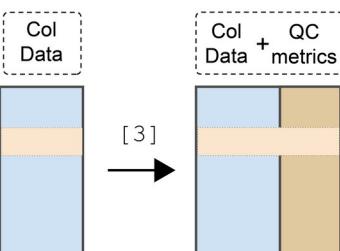
### Accessing Taxonomic Info.



[1] mia::addTaxonomyTree(tse)

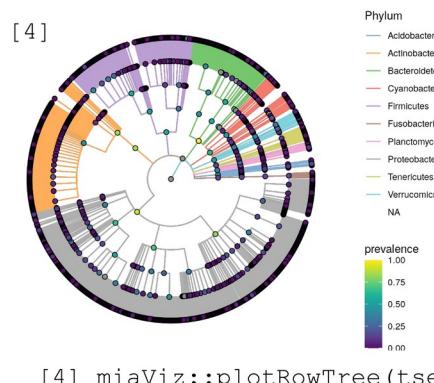
[2] TreeSE::aggValue(tse)

### Quality Control



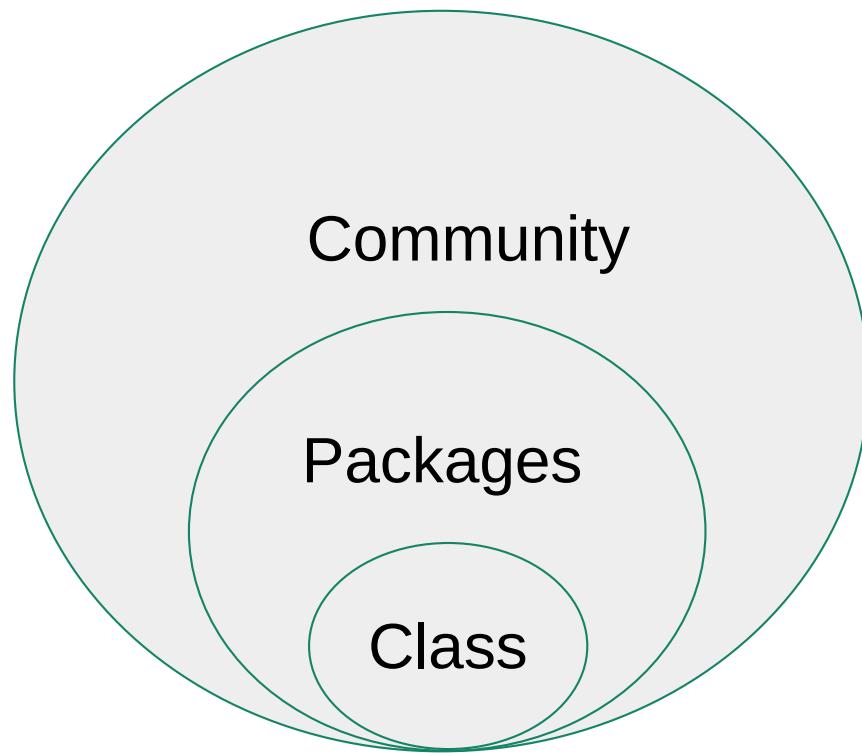
[3] scatter::addPerCellQC(tse)

### Visualizing with miaViz

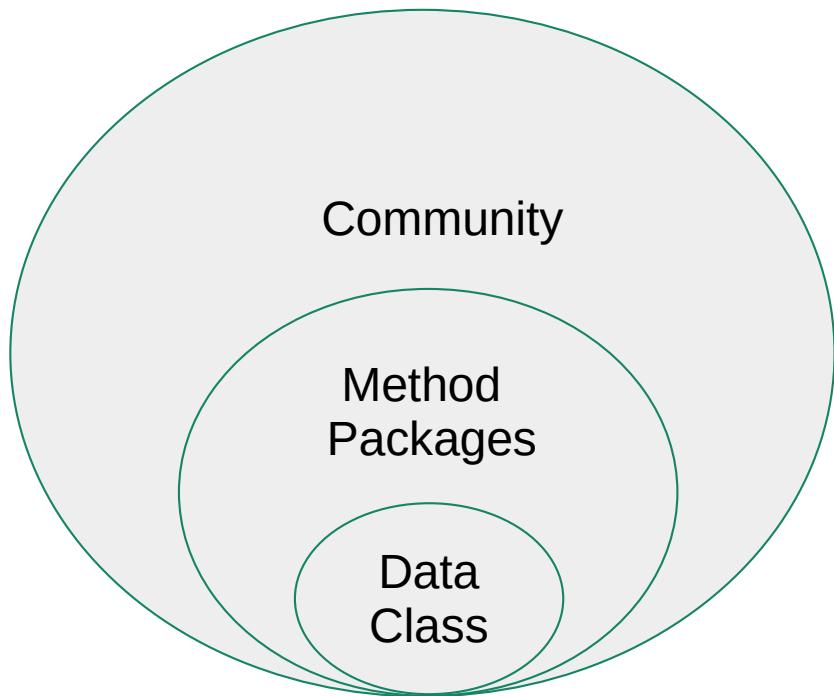


[4] miaViz::plotRowTree(tse)

*Standardized data containers*  
are central for the R/Bioc ecosystem



Reduce overlapping efforts, improve interoperability, ensure sustainability.



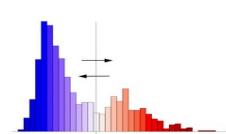
## Data packages

ExperimentHub

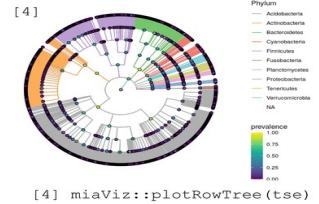
platforms all rank 76 / 1974 posts 2 / 1 / 2e+01 / 1 in Bioc 4 years  
build ok updated before release dependencies 72

DOI: [10.18129/B9.bioc.ExperimentHub](https://doi.org/10.18129/B9.bioc.ExperimentHub) [f](#) [t](#)

**mia – microbiome analysis**  
getDiversity(x)  
calculateDMM(x)



**miaViz - Visualization**



# Package ecosystem

Transparency

Reproducibility

Collaboration

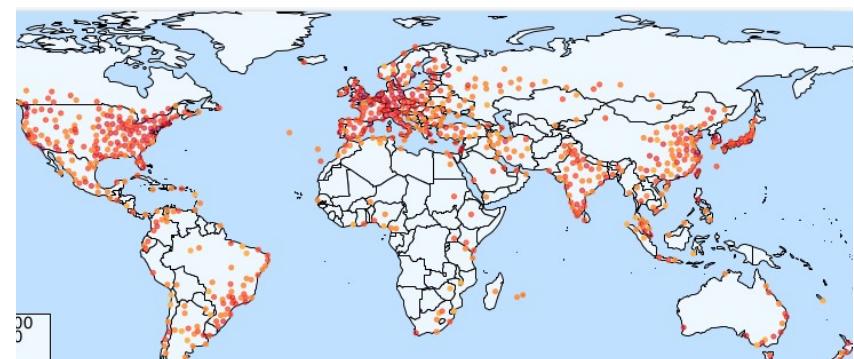
# R/Bioc community

(Global) Bioconductor conference

Asian Bioconductor conference

European Bioconductor conference

- *Heidelberg, Sep 14-16, 2022*



A screenshot of the EuroBioC 2022 conference website. The header includes the Bioconductor logo and navigation links for HOME, SUBMISSIONS, REGISTRATION, SCHEDULE, SPONSORS, and ABOUT. The main title is "EUROBIOC 2022 CONFERENCE, HEIDELBERG, BIOQUANT, SEPTEMBER 14-16, 2022." Below it is the tagline "WHERE SOFTWARE AND BIOLOGY CONNECT". In the bottom right corner, there's a decorative hexagonal graphic featuring a landscape with a castle and a musical note, with the text "EuroBioC2022" and "Heidelberg" and the website "www.bioconductor.org".

# Initializing reproducible report

**Task:** create Rmarkdown document that imports the data

## 1 Welcome!

1.1 Installation Instructions

2 Amazing Resources

3 Dynamic Documents

3.1 Reproducible Research

4 Markdown

5 RMarkdown

5.1 Why R Markdown?

5.2 Simple Workflow

5.3 Creating a .Rmd File

5.4 YAML Headers

5.5 Markdown Basics

# Creating Dynamic Documents with RMarkdown and Knitr

Code ▾

*By: Marian L. Schmidt, @micro\_marian , marschmi at umich.edu*

*May 11th, 2016*

## 1 Welcome!

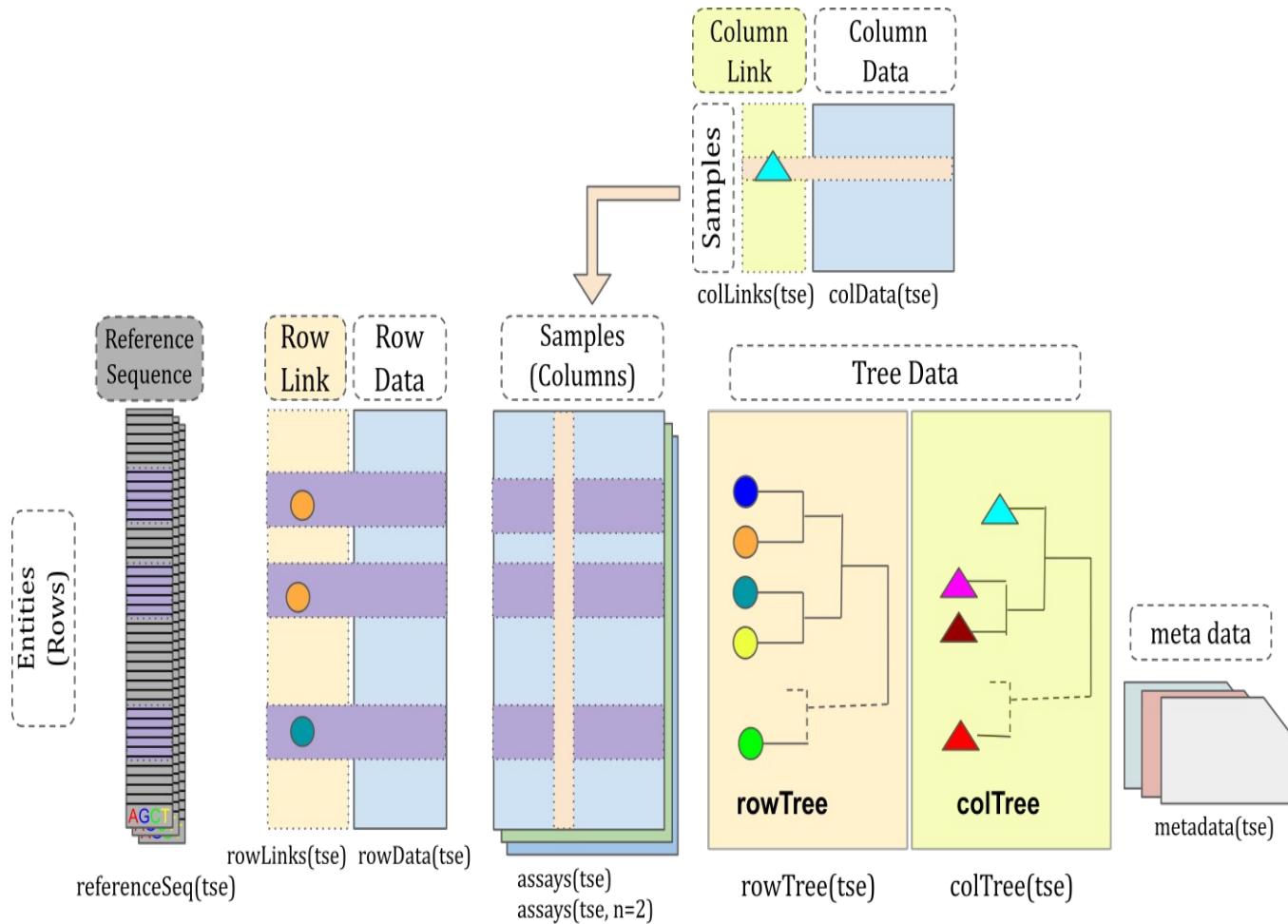
This workshop was hosted on **May 11th, 2016** and has 2 supplemental materials available:

1. **Class notes** can be found on the workshop [etherpad](#).
2. This workshop was **recorded** live and is available on [YouTube](#). Please be welcome to tune in on YouTube!

This tutorial was constructed as a part of Dr. C Titus Brown's [Data Intensive Biology \(DIB\)](#) training program at the University of California, Davis. The DIB training program hosts local, remote workshops, summer schools, and

# Importing microbiome data to R

**Task:** import biom data files for your chosen case study into a specific data container (structure) in R, TreeSummarizedExperiment (TSE) Huang et al. (2020). This provides the basis for downstream data analysis.

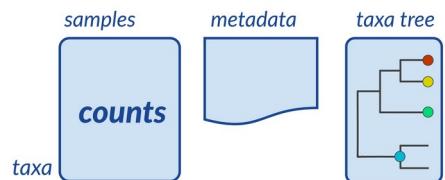


# Data containers support collaborative development of analysis methods & workflows

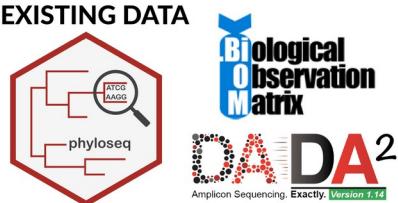
## Import Data

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

### RAW DATA

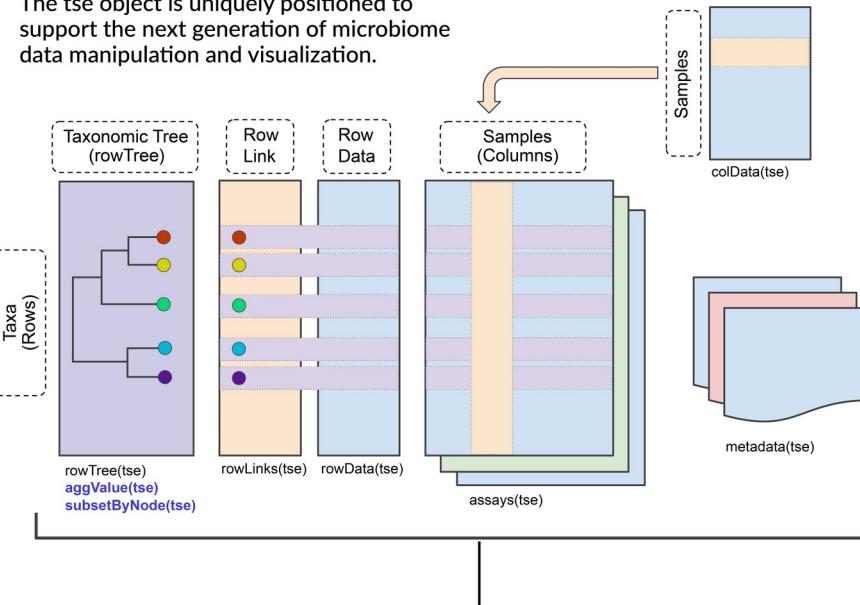


### EXISTING DATA



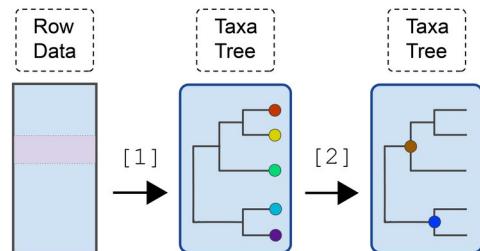
## The TreeSE object

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



## The mia Pipeline

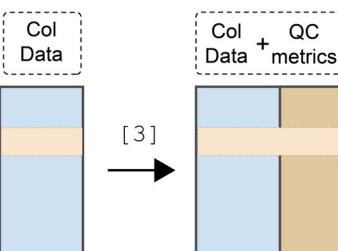
### Accessing Taxonomic Info.



[1] mia::addTaxonomyTree(tse)

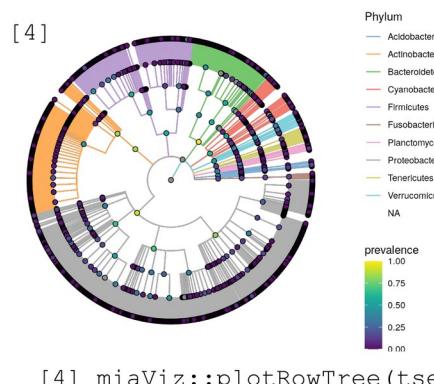
[2] TreeSE::aggValue(tse)

### Quality Control



[3] scatter::addPerCellQC(tse)

### Visualizing with miaViz



[4] miaViz::plotRowTree(tse)

# Orchestrating Microbiome Analysis with R and Bioconductor – online book: *beta version*

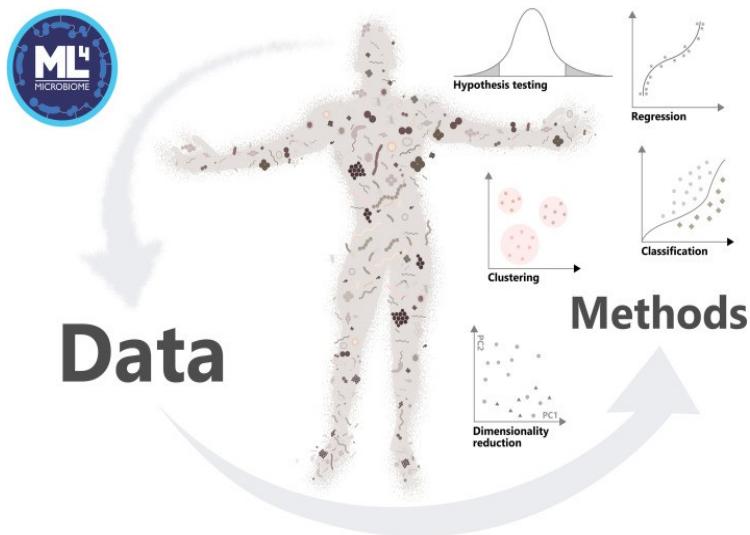


Figure source: Moreno-Indias et al. (2021) Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. *Frontiers in Microbiology* 12:11.

## Microbiome Analysis

**Authors:** Leo Lahti [aut], Sudarshan Shetty [aut], Felix GM Ernst [aut, cre]

**Version:** 0.98.0003

**Modified:** 2020-12-06

**Compiled:** 2020-12-13

**Environment:** R version 4.0.0 (2020-04-24), Bioconductor 3.11

**License:** CC BY-NC-SA 3.0 US

**Copyright:**

**Source:** <https://github.com/microbiome/MiaBook>

## Preface

This website is a book on microbiome analysis in the Bioconductor universe and is showing common principles and workflows of performing microbiome analysis.

The book was borne out of necessity, while updating tools for microbiome analysis to work with common classes of the Bioconductor project handling count data of various sorts. It is heavily influenced by similar resources, such as the [Orchestrating Single-Cell Analysis with Bioconductor](#) book, [phyloseq tutorials](#) and [microbiome tutorials](#).

We focus on microbiome analysis tools, new, updated and established methods. In the *Introduction* section, we show how to work with the key data infrastructure `TreeSummarizedExperiment` and related classes, how this framework relates to other infrastructure and how to load microbiome analysis data to work with in the context of this framework.

The second section, *Focus Topics*, is all about the steps for analyzing microbiome data, beginning with the most common steps and progressing to more specialized methods in subsequent sections.

The third section, *Appendix*, contains the rest of things we didn't find another place for, yet.

Home

CC BY-NC-SA

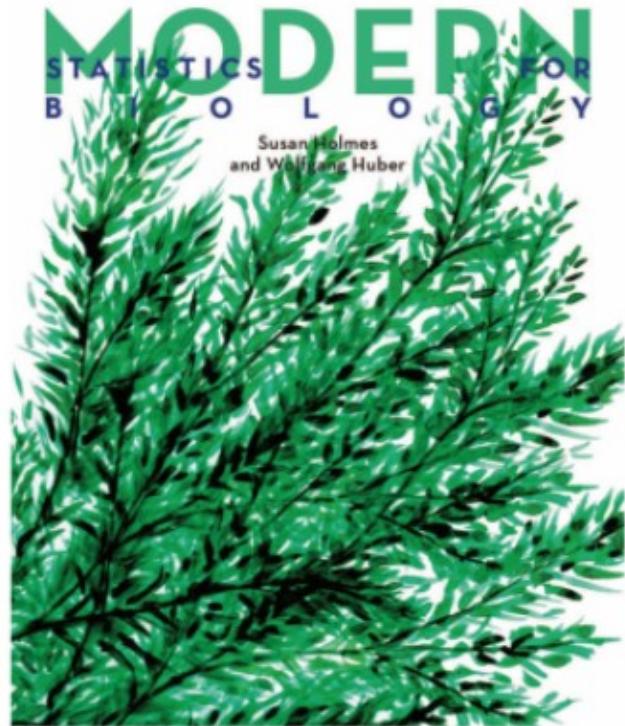


Figure 5: The online version provides the text in HTML, data files and up-to-date code.

---

[1 Generative Models for Discrete Data](#)

[2 Statistical Modeling](#)

[3 High-Quality Graphics in R](#)

[4 Mixture Models](#)

[5 Clustering](#)

[6 Testing](#)

[7 Multivariate Analysis](#)

[8 High-Throughput Count Data](#)

[9 Multivariate Methods for Heterogeneous Data](#)

[10 Networks and Trees](#)

[11 Image Data](#)

[12 Supervised Learning](#)

[13 Design of High-Throughput Experiments and Their Analyses](#)