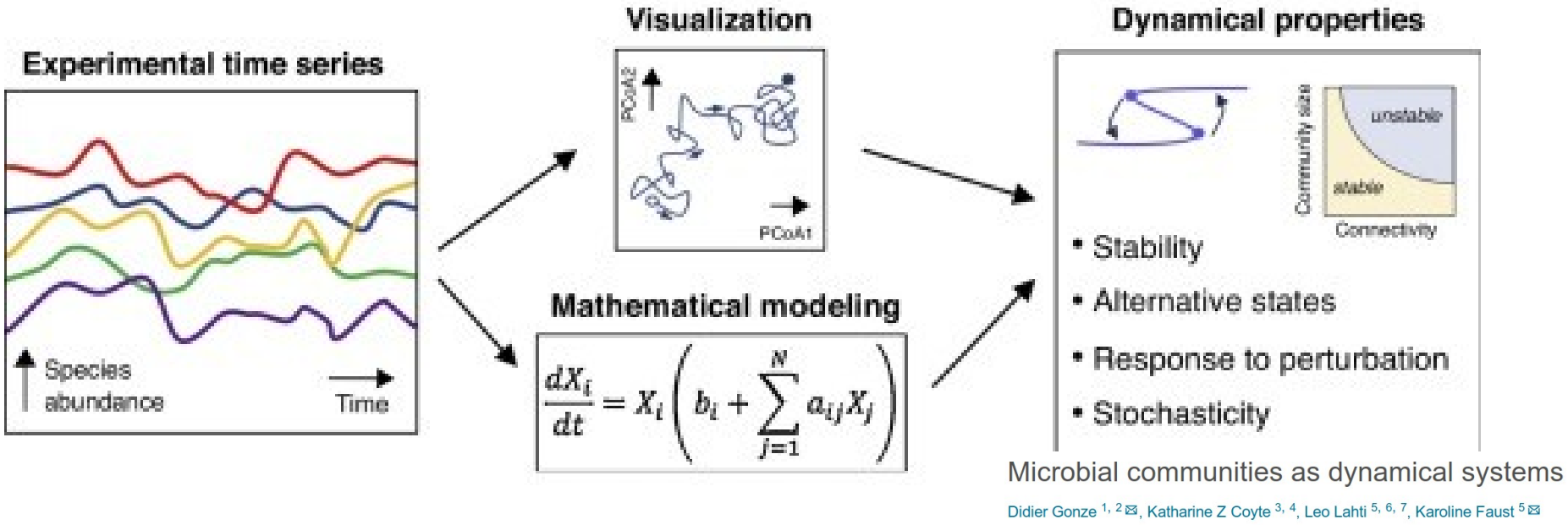
# Advances in Microbiome Data Science with R/Bioconductor

Workshop, Tirana May 24, 2022




Microbial communities as dynamical systems

Didier Gonze [1, 2]✉, Katharine Z Coyte [3, 4], Leo Lahti [5, 6, 7], Karoline Faust [5]✉

**Associate Prof. Leo Lahti** | datascience.utu.fi
Department of Computing, University of Turku, Finland

Microbiome bioinformatics

Data science workflows

Open science & reproducible research

The demise of alchemy provides further evidence, if further evidence were needed, that 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*. Alchemy, as a clandestine enterprise, could never develop a community of the right sort. Popper was right to think that science can flourish only in an open society.

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

Open
reporting and communication
were part of academic culture
since the early days

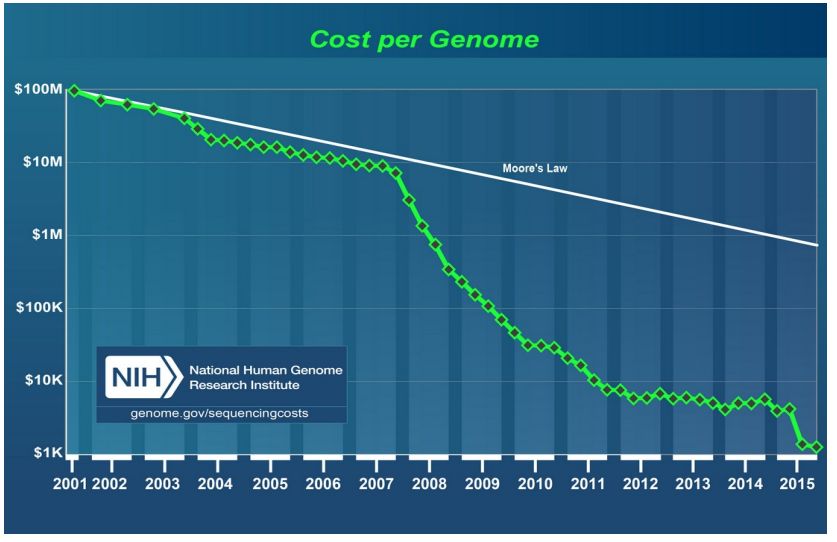
A family of alchemists at work, an engraving by Philip Galle, after a painting by Pieter Bruegel the Elder, published by Hieronymus Cock, *c*.1558.

# Initial sequencing and analysis of the human genome

~2001

**International Human Genome Sequencing Consortium***

** A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

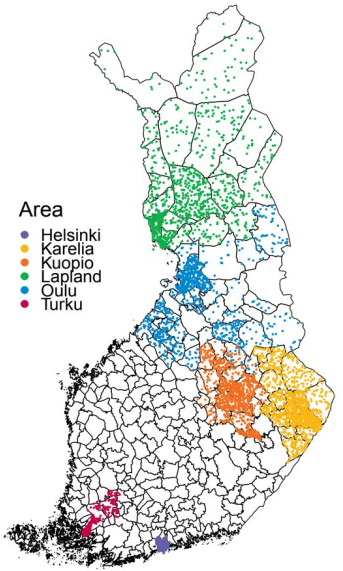



The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. **Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome**. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

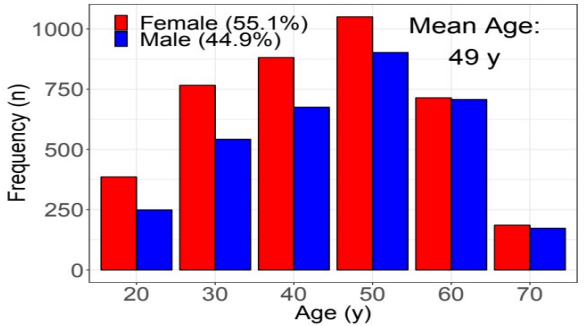

# FINRISK cohort (2002)

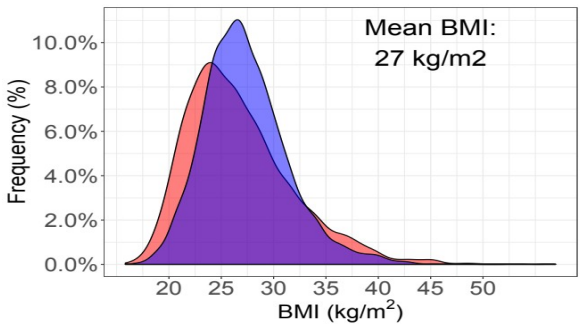
20 year follow-up

Population register data


NATIONAL INSTITUTE FOR HEALTH AND WELFARE



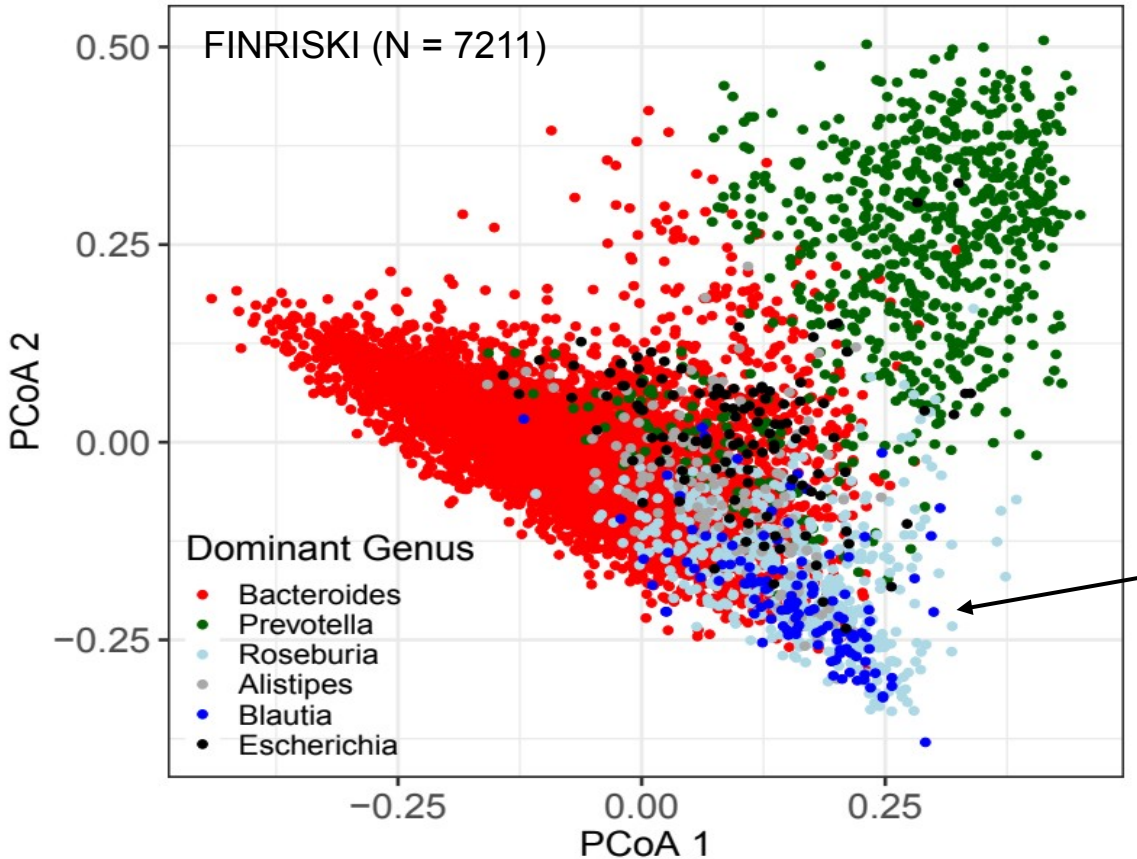



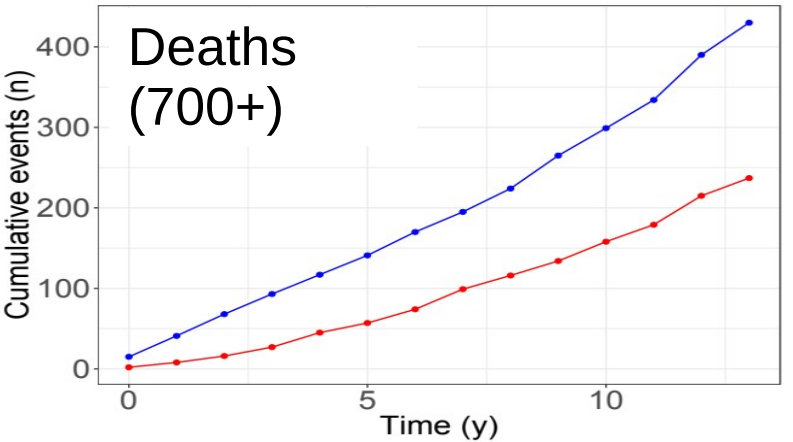

**Taxonomic signatures of cause-specific mortality risk in human gut microbiome**

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

*Nature Communications* **12**, Article number: 2671 (2021) | Cite this article

# Gut microbiome and mortality risk




You?

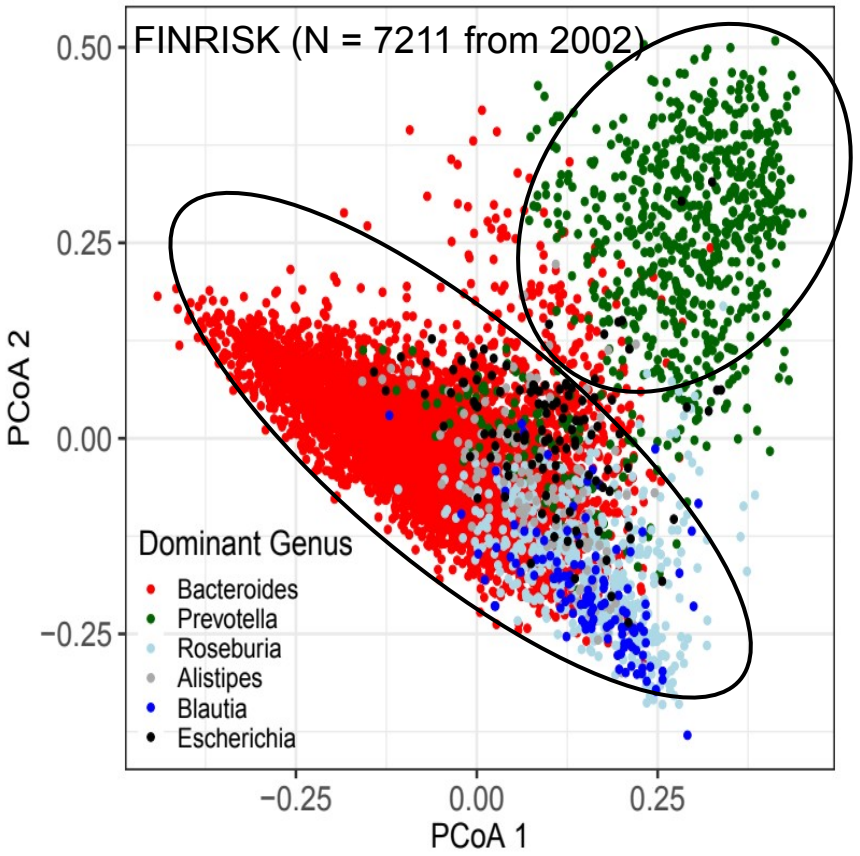
# There's A New Self-Administered Medical Craze For 2018 And It's Much More Horrifying Than Tide Pods

1 SHARES

Share on Facebook | Share on Twitter | +

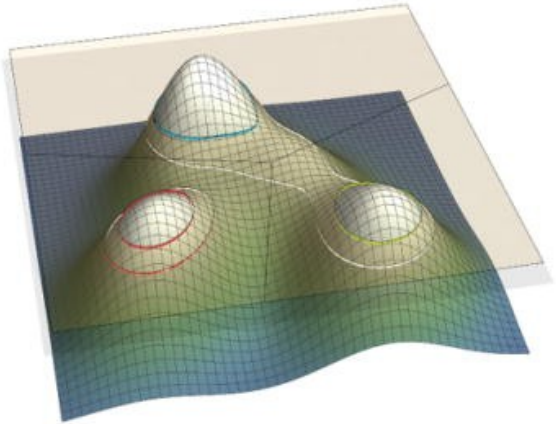
**People are attempting DIY fecal transplants without medical supervision and putting themselves at risk of deadly infections, warn experts**

# Gut microbiome similarity among Finnish adults




- How many clusters?
- Which clusters?
- Where are the boundaries
- Who belongs to which cluster?

# How to choose a correct model?
# → a community typing example



Enterotypes in the landscape of gut microbial community composition. Costea *et al*. Nature 2018.

$$2 \times 6^6 = 93312$$

**Taxonomic level**
- Phylum
- Family
- Order
- Genus
- Species
- Strain..

**Filtering**
- None
- Prevalent
- Core
- Excl. outliers
- High variance
- Custom

**Normalization**
- None
- TSS
- CSS
- ILR/ALR/CLR
- phILR
- Hellinger

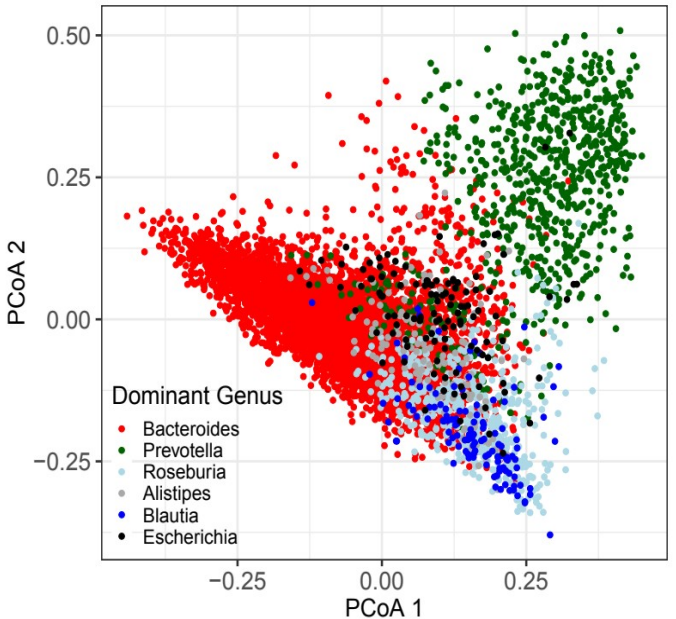
**(Dis)similarity**
- Eulidean
- Aitchison
- Bray-Curtis
- Jaccard
- weighted Unifrac
- unweighted Unifrac

**Clustering method**
- Hierarchical / Ward
- Hierarchical / Complete
- Gaussian mixture
- DMM
- PAMR
- K-means

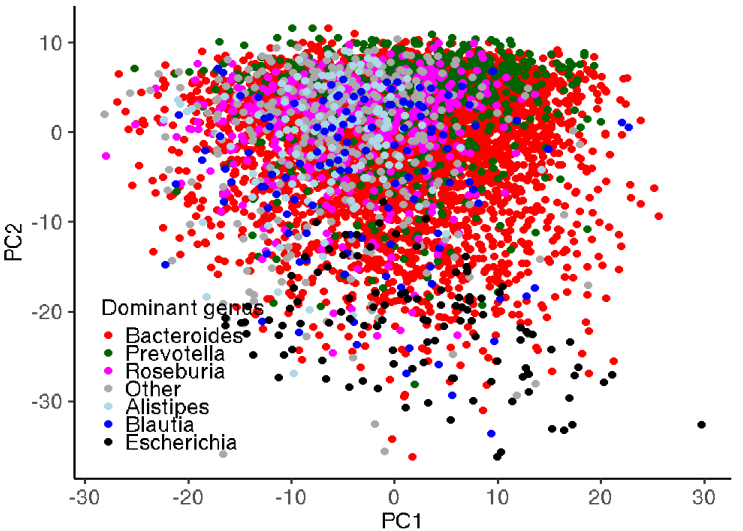
**Regulation**
- Calinski-Harabasz
- Dirichlet Process
- Silhouette Index
- AIC
- BIC
- DIC

Walk-through example in R/Bioc by Holmes & McMurdie
http://statweb.stanford.edu/~susan/papers/EnterotypeRR.html

# PCoA + Bray-Curtis




# PCA + Aitchison




## Reproducible Research: Enterotype Example

*Susan Holmes and Joey McMurdie*

http://statweb.stanford.edu/~susan/papers/EnterotypeRR.html

Comment on this paper

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

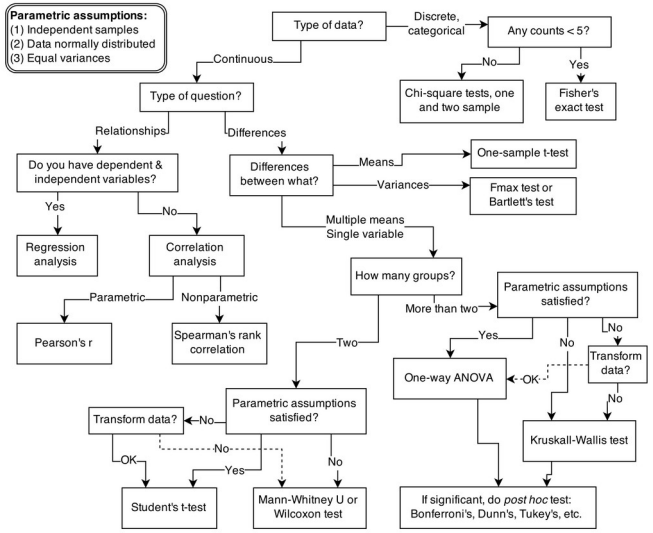
Aaro Salosensaari, Ville Laitinen, Aki Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, Georg Alfthan, Michael Inouye, Jeramie D. Watrous, Tao Long, Rodolfo Salido, Karenina Sanders, Caitriona Brennan, Gregory C. Humphrey, Jon G. Sanders, Mohit Jain, Pekka Jousilahti, Veikko Salomaa,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

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

# The influence of hidden researcher decisions in applied microeconomics

Nick Huntington-Klein ✉, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, Yaniv Stopnitzky

First published: 22 March 2021
https://doi.org/10.1111/ecin.12992




**Researchers make hundreds of decisions about data collection, preparation, and analysis in their research.** We use a many-analysts approach to measure the extent and impact of these decisions. Two published causal empirical results are replicated by seven replicators each. *We find large differences in data preparation and analysis decisions, many of which would not likely be reported in a publication.* No two replicators reported the same sample size. *Statistical significance varied across replications, and for one of the studies the effect's sign varied as well.* The standard deviation of estimates across replications was 3–4 times the mean reported standard error.

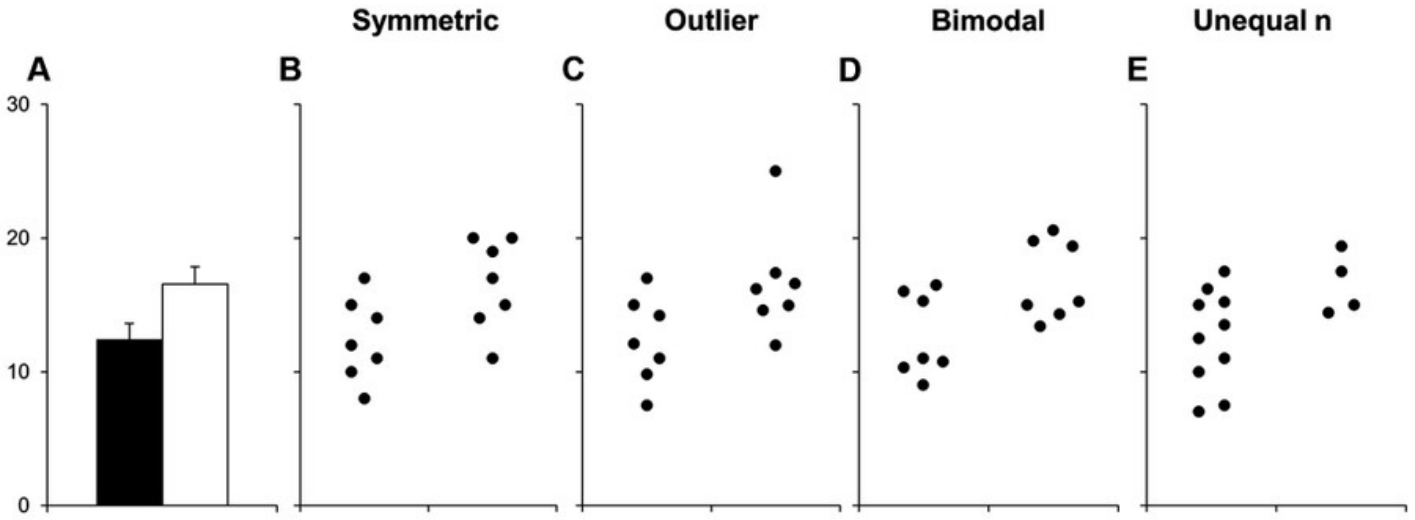
P < 0.04

Effect?

# P < 0.05

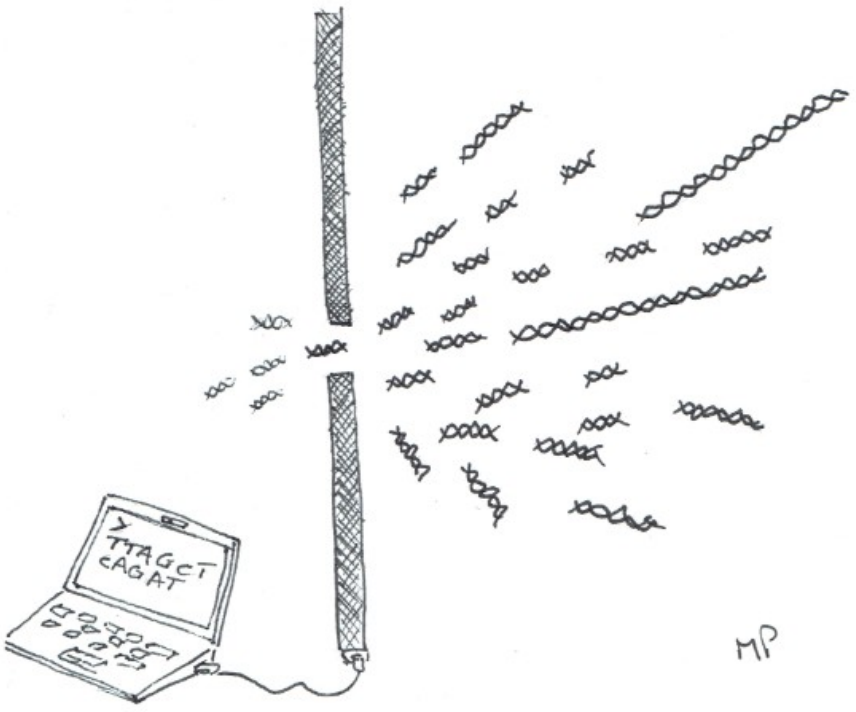
P < 0.06

No effect?

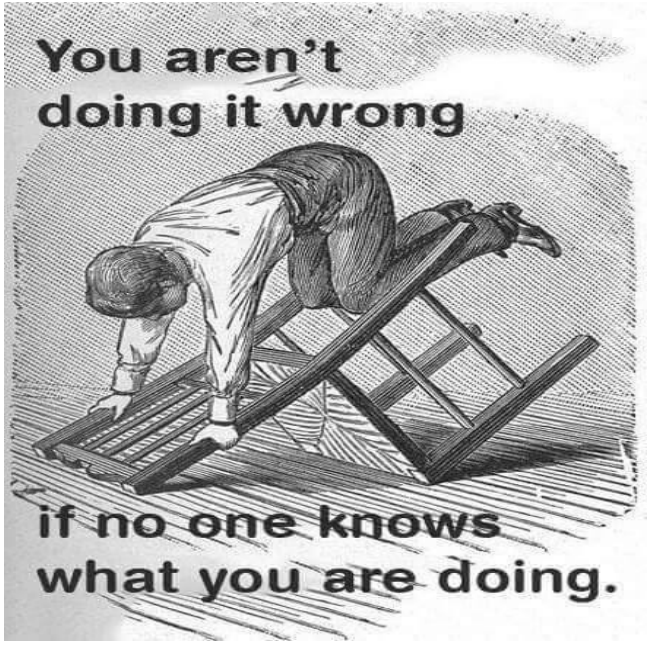



| Test | p value | | | |
|---|---|---|---|---|
| T-test: Equal var. | 0.035 | 0.050 | 0.026 | 0.063 |
| T-test: Unequal var. | 0.035 | 0.050 | 0.026 | 0.035 |
| Wilcoxon | 0.054 | 0.073 | 0.128 | 0.103 |

(barely) not statistically significant (p=0.052)
a barely detectable statistically significant difference (p=0.073)
a borderline significant trend (p=0.09)
a certain trend toward significance (p=0.08)
a clear tendency to significance (p=0.052)
a clear trend (p<0.09)
a clear, strong trend (p=0.09)
a considerable trend toward significance (p=0.069)
a decreasing trend (p=0.09)
a definite trend (p=0.08)
a distinct trend toward significance (p=0.07)
a favorable trend (p=0.09)
a favourable statistical trend (p=0.09)
a little significant (p<0.1)
a margin at the edge of significance (p=0.0608)
a marginal trend (p=0.09)
a marginal trend toward significance (p=0.052)
a marked trend (p=0.07)
a mild trend (p<0.09)
a moderate trend toward significance (p=0.068)
a near-significant trend (p=0.07)
a negative trend (p=0.09)
a nonsignificant trend (p<0.1)
a nonsignificant trend toward significance (p=0.1)



?


RESEARCH PRIORITIES

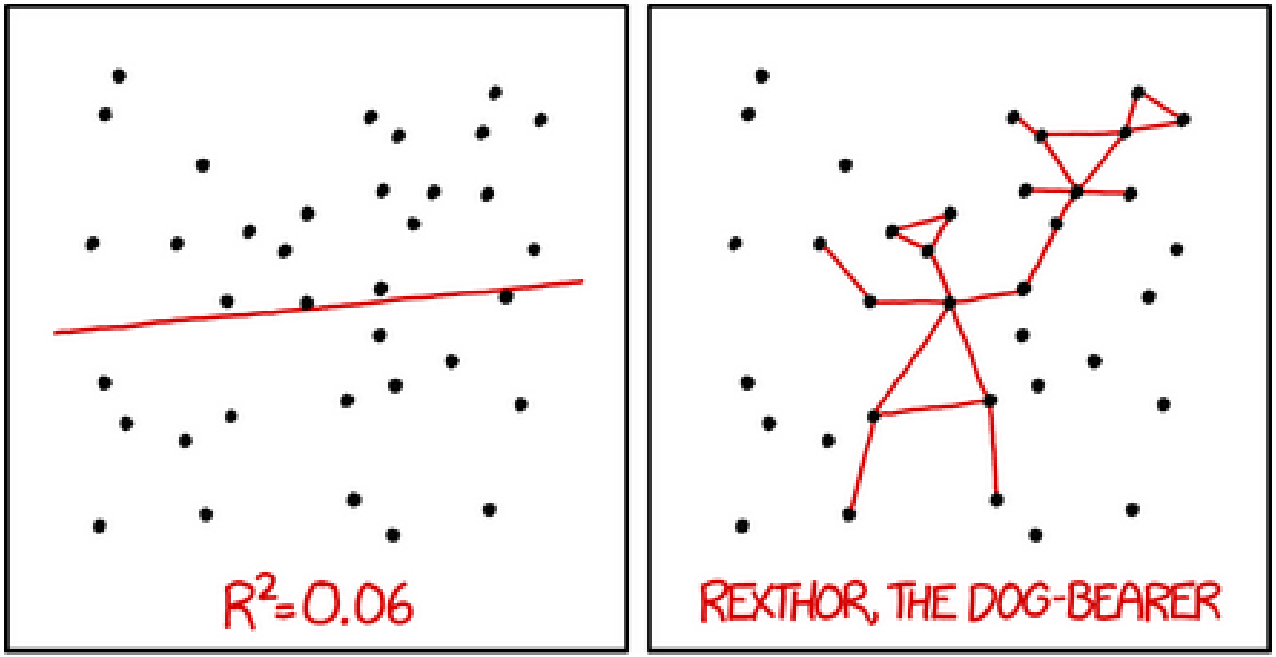
**Shining Light into Black Boxes**

A. Morin[1], J. Urban[2], P. D. Adams[3], I. Foster[4], A. Sali[5], D. Baker[6], P. Sliz[1,*]

```
int getRandomNumber()
{
    return 4;  // chosen by fair dice roll.
               // guaranteed to be random.
}
```

http://web.stanford.edu/class/cs109l/unrestricted/images/

# How we choose which model to apply?

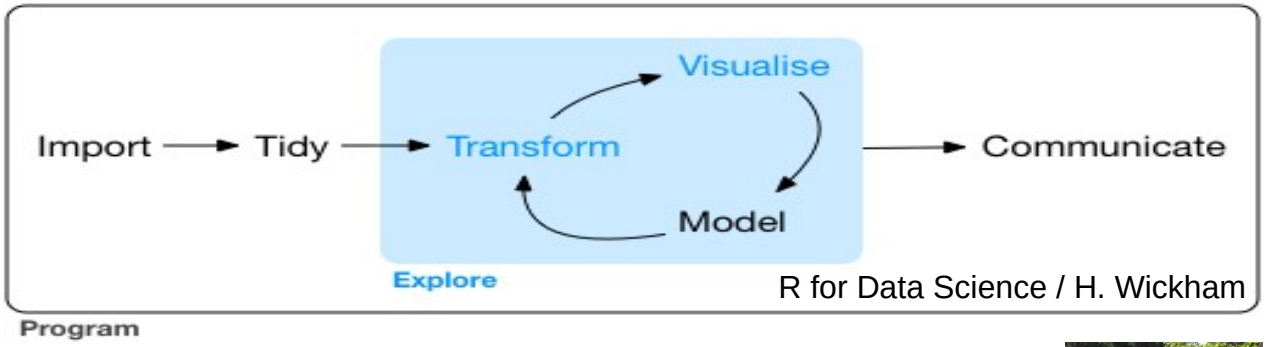


# Data science workflow

## reproducibility & automation

## sharing & collaboration

## transparency

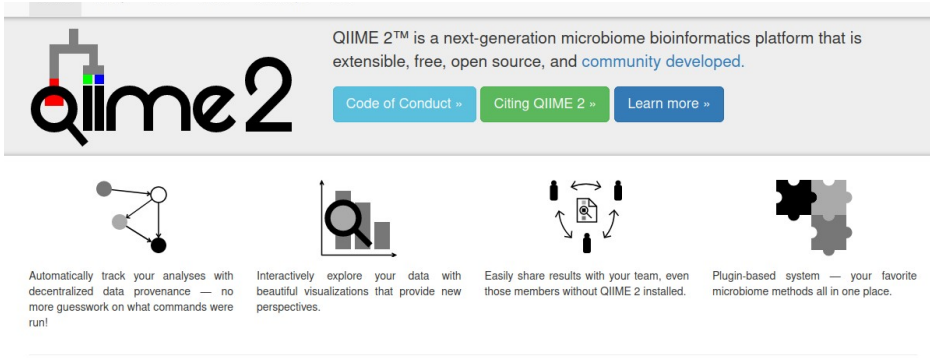



R for Data Science / H. Wickham

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

Ben J. Callahan[1], Kris Sankaran[1], Julia A. Fukuyama[1], Paul J. McMurdie[2], Susan P. Holmes

# open data science ecosystems


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

PeerJ >

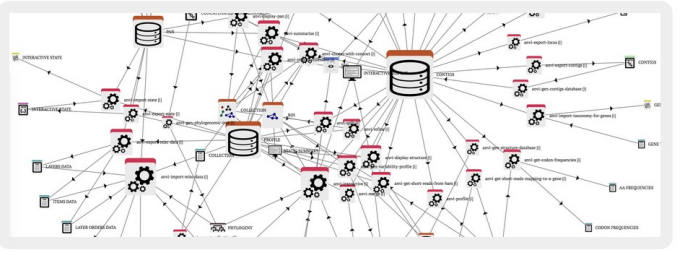
## Anvi'o: an advanced analysis and visualization platform for 'omics data

Research article | Bioinformatics | Biotechnology | Computational Biology | Genomics | Microbiology

A. Murat Eren [1,2], Özcan C. Esen [1], Christopher Quince [3], Joseph H. Vineis [1], Hilary G. Morrison [1], Mitchell L. Sogin [1], Tom O. Delmont [1]

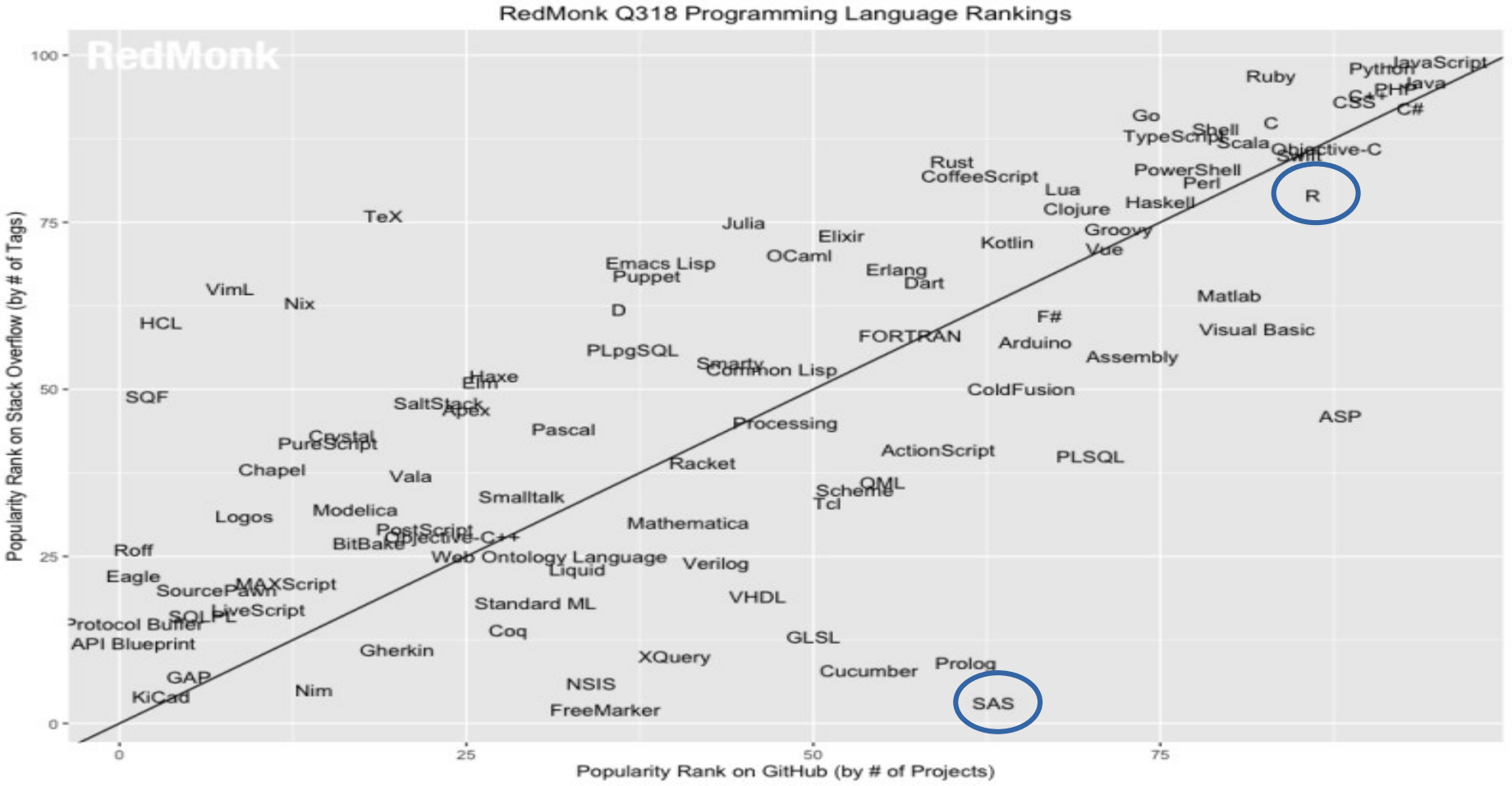
Published October 8, 2015

## Anvi'o in a nutshell



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

# Varying cultures of open collaboration



# What is Bioconductor ?

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

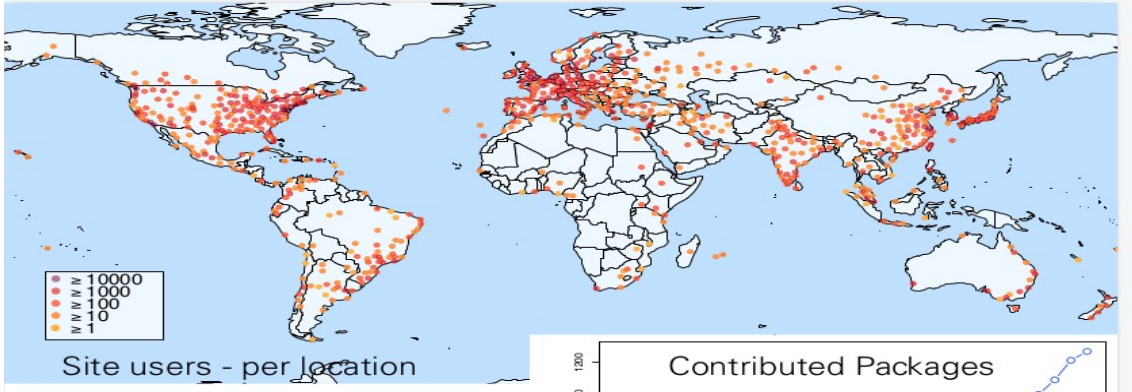
Started 2001 as a platform for analysis & understanding of microarray data

More than 1,600 packages. Domains of expertise:

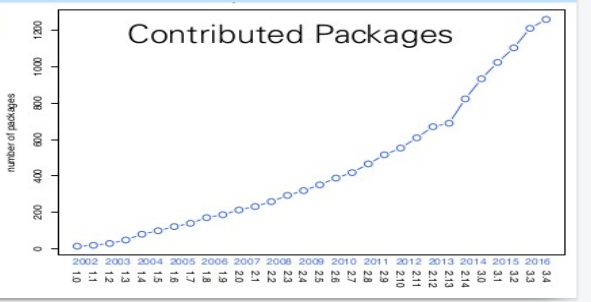
- Sequencing (RNASeq, ChIPSeq, single-cell, called variants, ...)
- Microarrays (methylation, expression, copy number, ...)
- Flow cytometry
- Proteomics
- Multi-Omics data integration

Important themes

- Reproducible research
- Interoperability between packages & workflows
  … even from different authors
- Usability




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

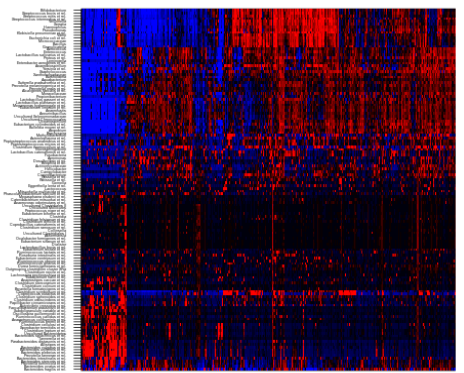


# ”Omics” data

## taxonomic abundance table

*Omics* in Oxford English Dictionary:
*in cellular and molecular biology*,
forming nouns with the sense
"all constituents considered collectively"

Features x samples

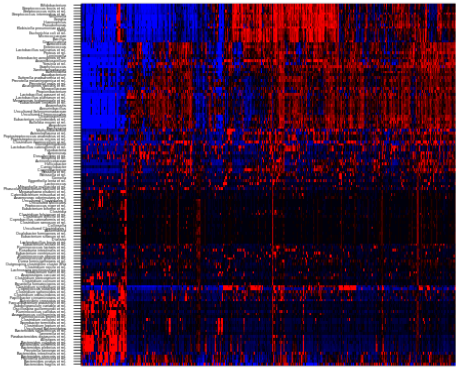


Genomics
Epigenomics
Microbiomics
Lipidomics
Proteomics
Glycomics
Foodomics
Transcriptomics
Metabolomics
Culturomics

Gut microbiota: 1000 western adults
(Lahti *et al.* Nature Comm. 2014)

Sample side information

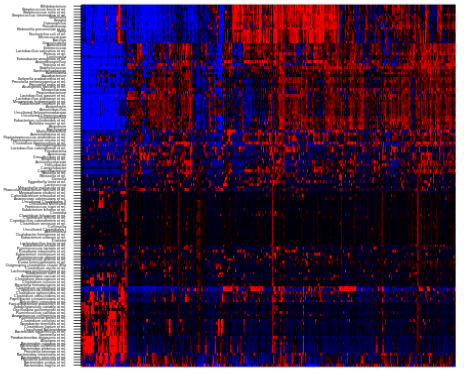
Features x samples

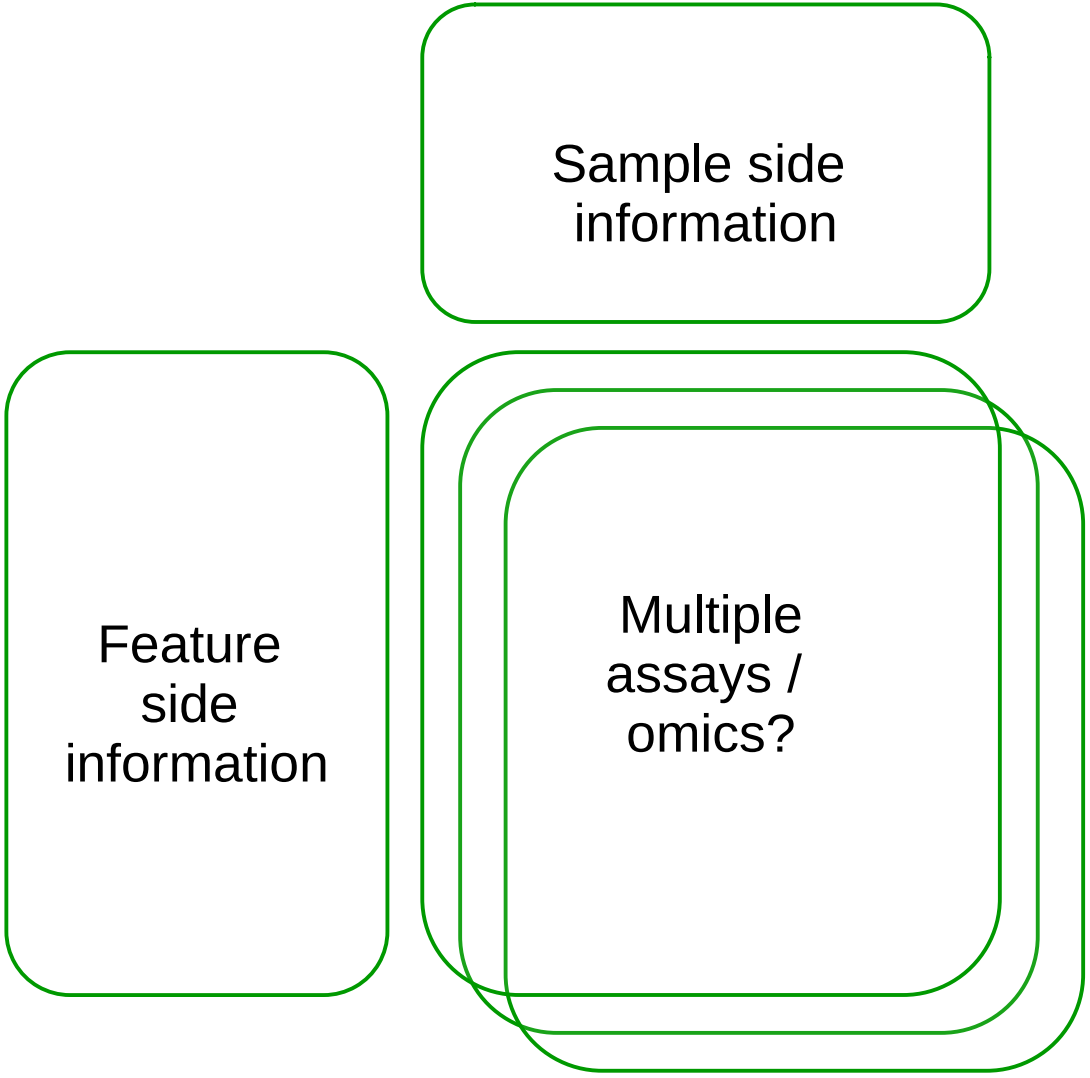


Sample side information

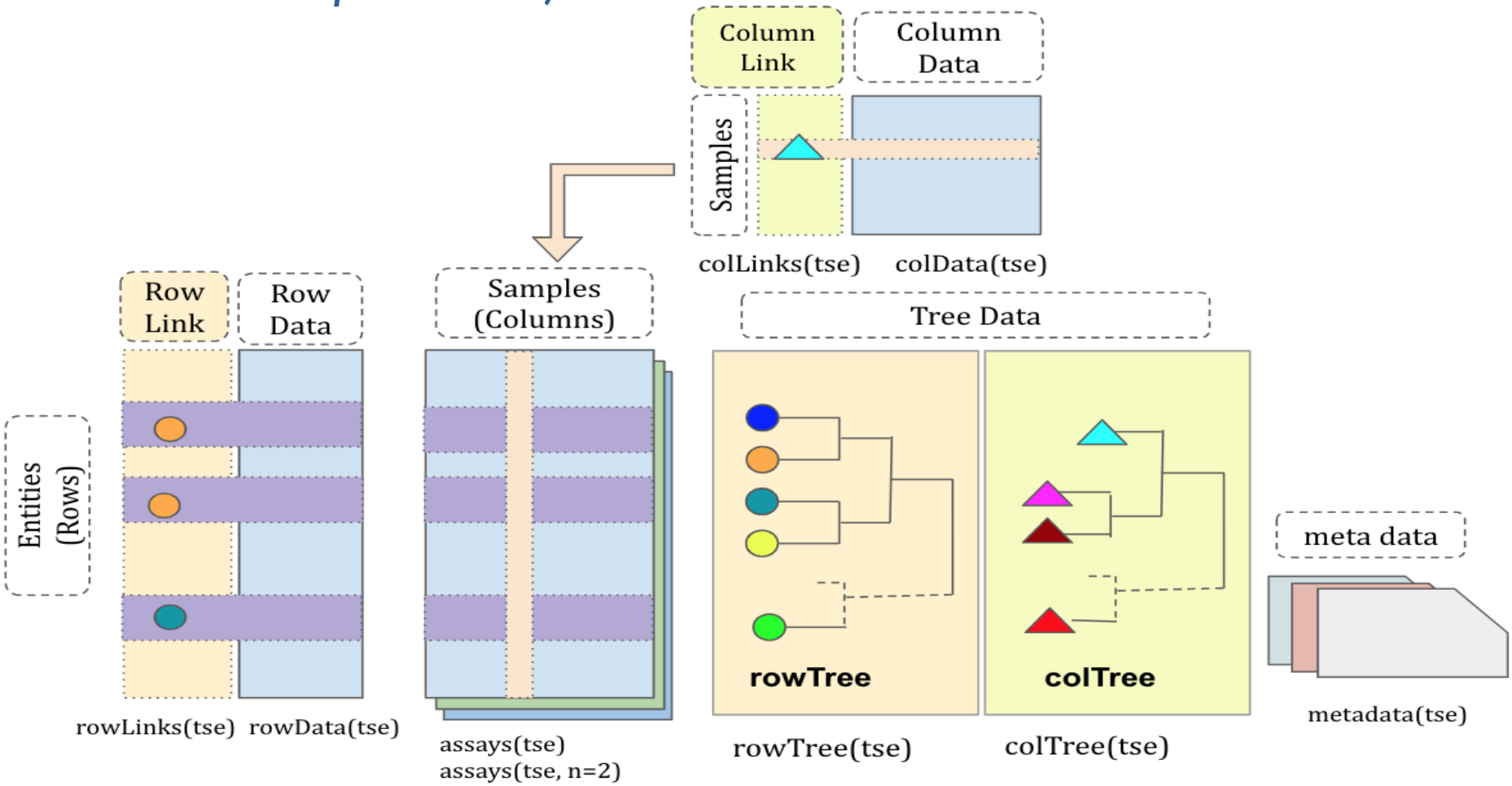
Feature side information

Features x samples

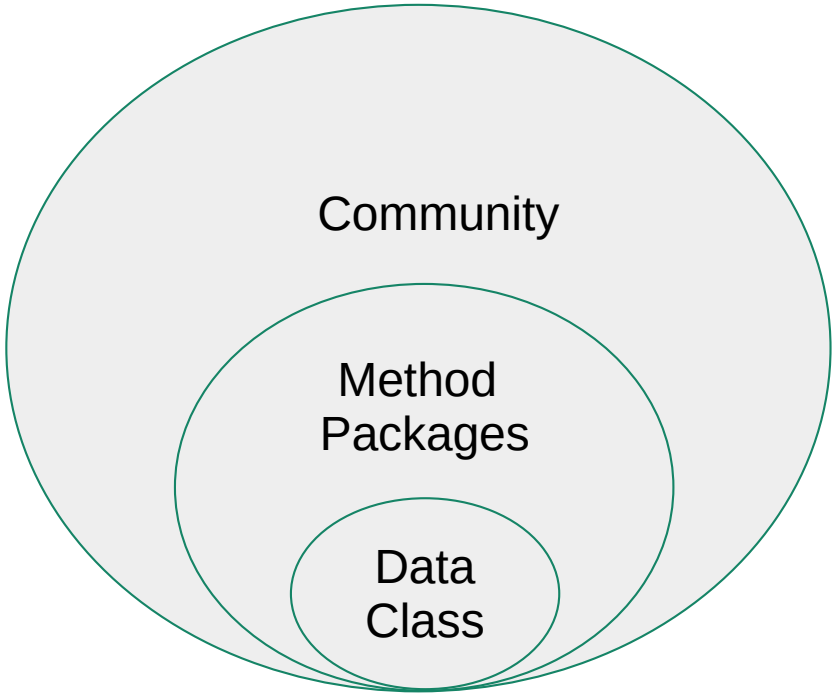



Sample side information
Feature side information
Multiple assays / omics?

# “Data container” (*TreeSummarizedExperiment*)



Reduce overlapping efforts, improve interoperability, ensure sustainability.




**Data packages**

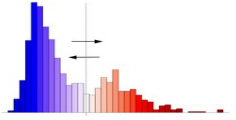
ExperimentHub

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

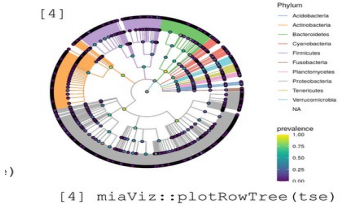
DOI: 10.18129/B9.bioc.ExperimentHub

**mia – microbiome analysis**

getDiversity(x)
calculateDMM(x)



**miaViz - Visualization**




# Package ecosystem

```
pheatmap(mat, annotation_row = taxa_clusters,
         annotation_col = sample_data,
         breaks = breaks,
         color = colors)
```

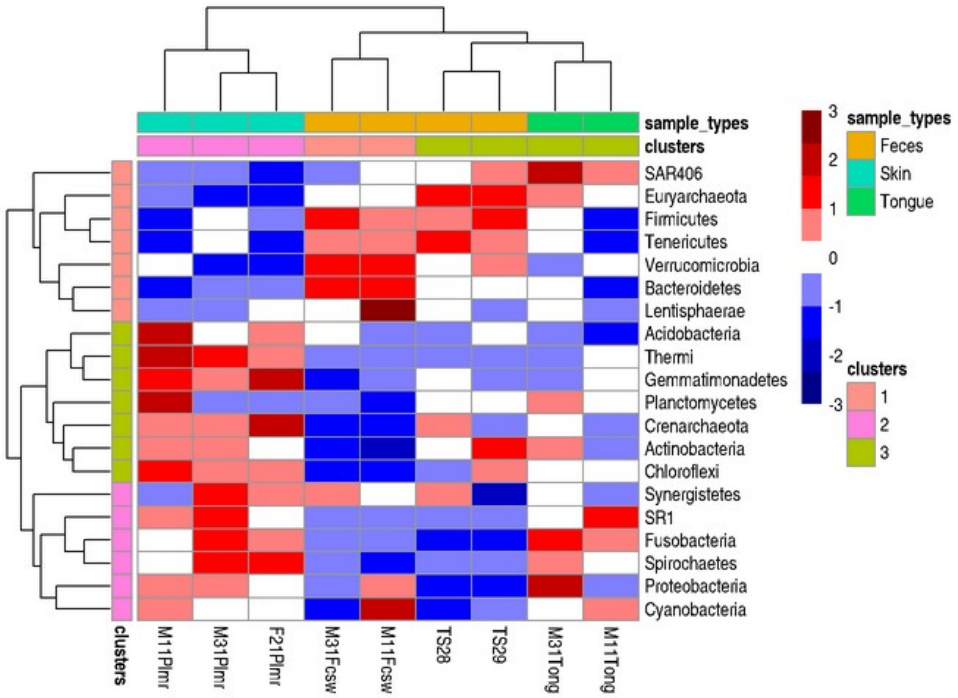
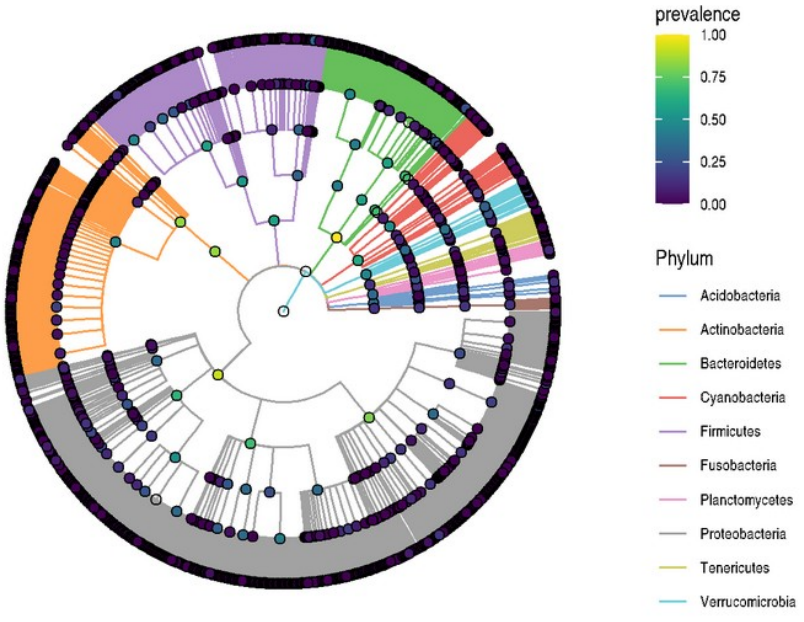
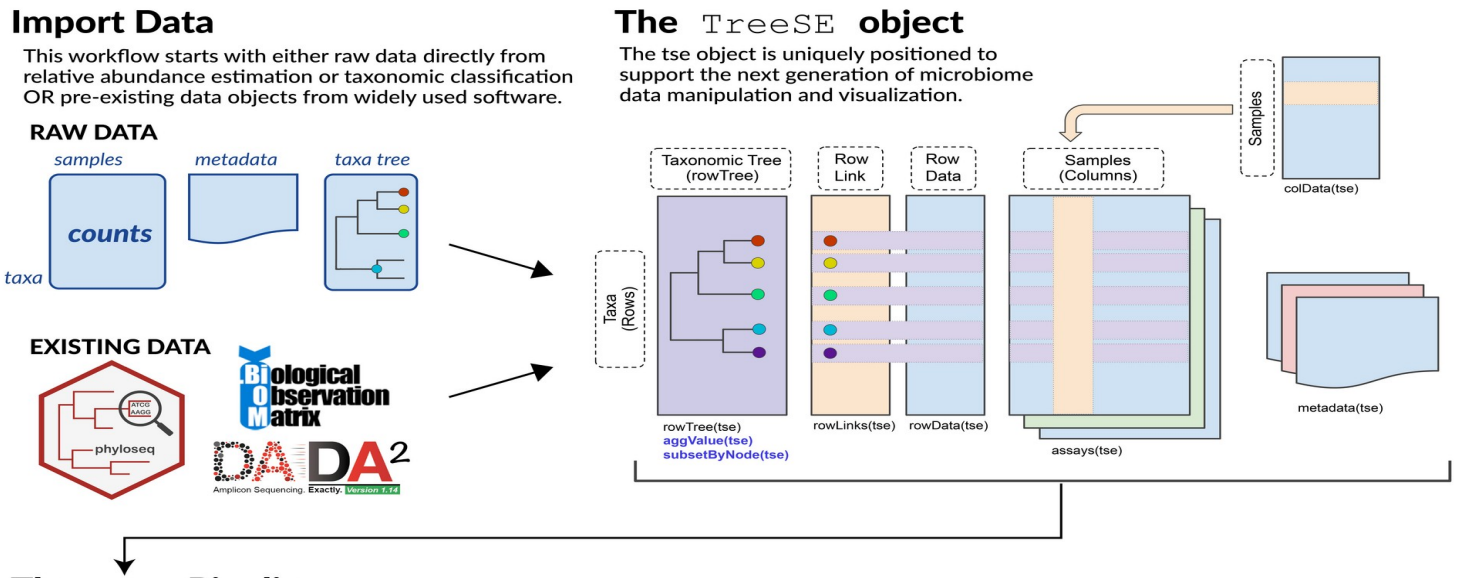



Figure 7.1: Prevalence of top phyla as judged by prevalence

```
plotRowTree(x[rowData(x)$Phylum %in% top_phyla_mean,],
            edge_colour_by = "Phylum",
            tip_colour_by = "prevalence",
            node_colour_by = "prevalence")
```

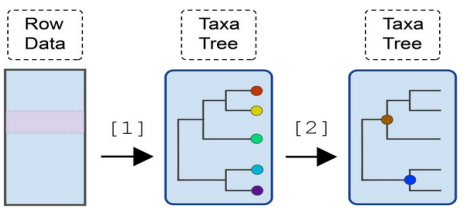


# Open & reproducible workflow



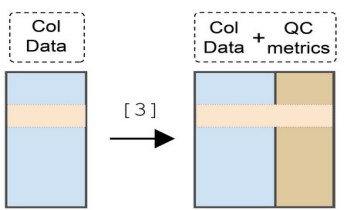

## The `mia` Pipeline

### Accessing Taxonomic Info.




```
[1] mia::addTaxonomyTree(tse)
[2] TreeSE::aggValue(tse)
```

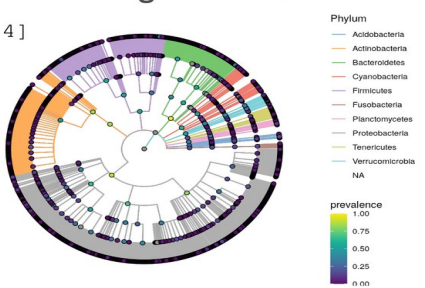
### Quality Control



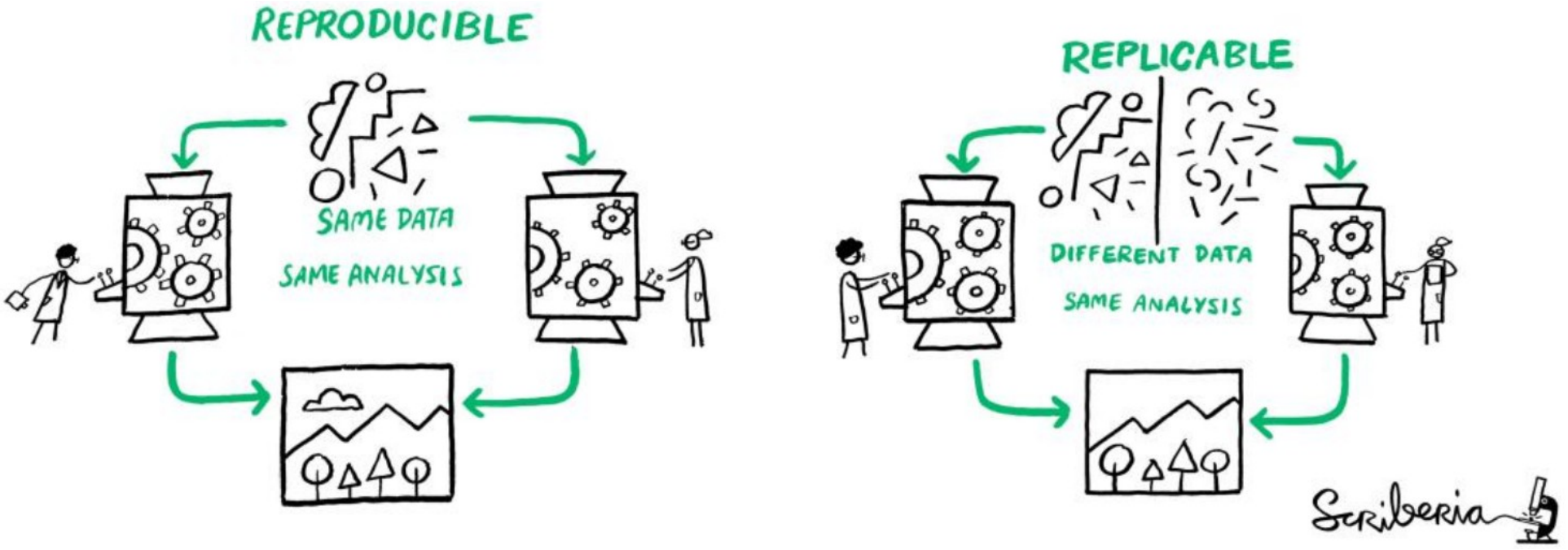

```
[3] scatter::addPerCellQC(tse)
```

### Visualizing with `miaViz`

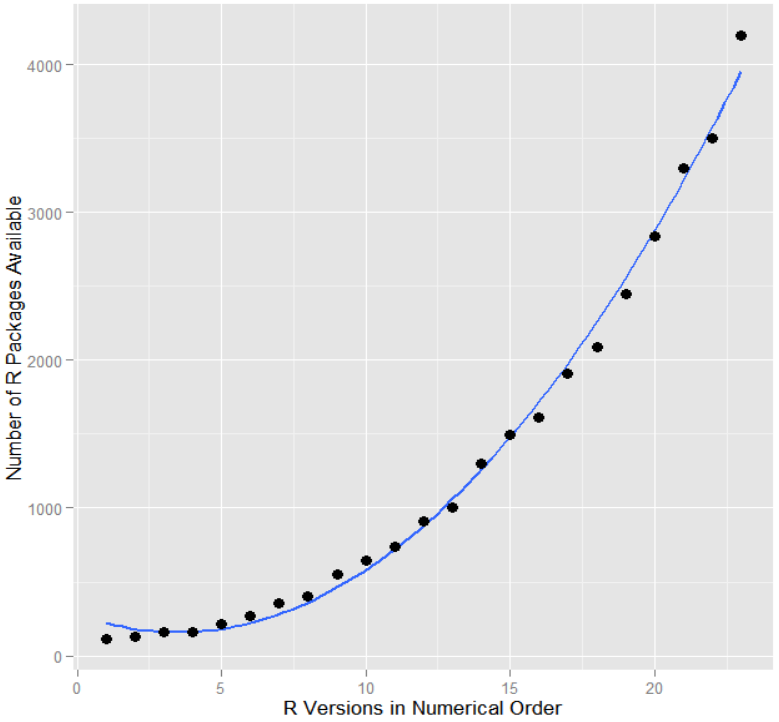
[4]




```
[4] miaViz::plotRowTree(tse)
```



# Number of open analysis tools has grown exponentially




github
SOCIAL CODING

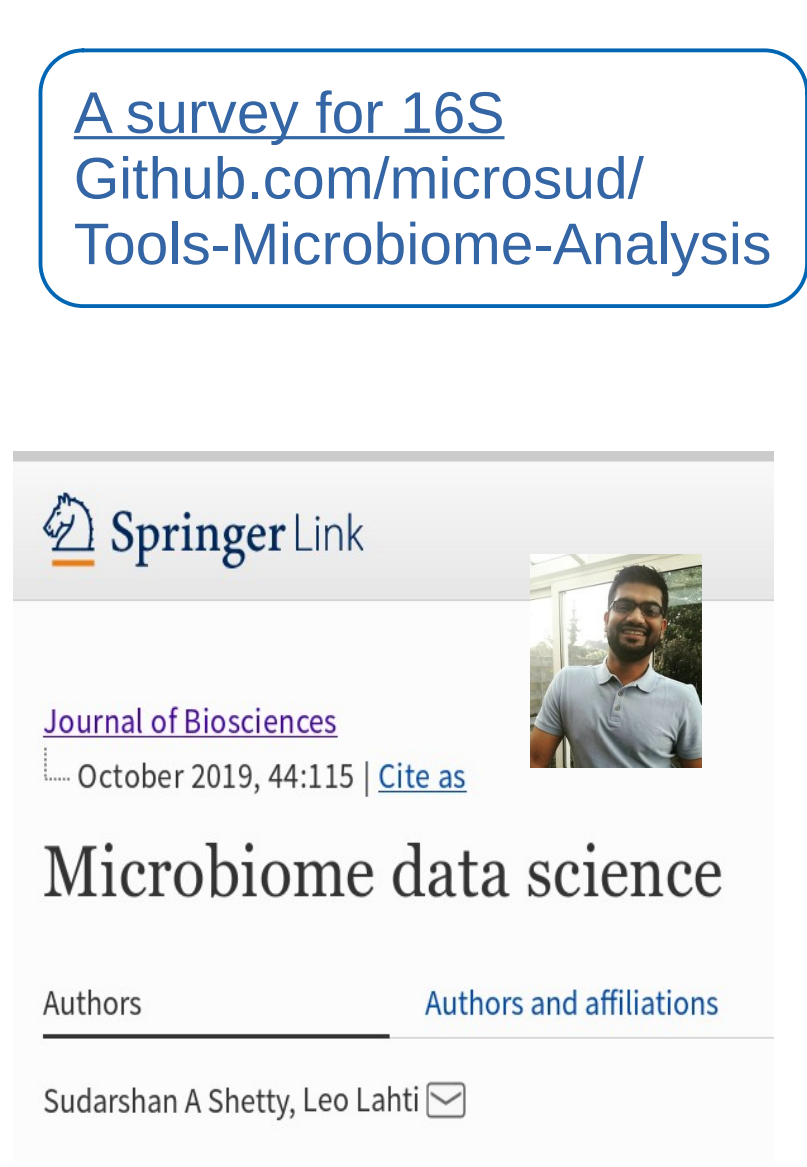
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

R

ROpenSci

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. MMinte Methodology for the large-scale assessment of microbial metabolic interactions (MMinte) 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. qiimer R tools compliment qiime
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



# Orchestrating Microbiome Analysis

***Authors:*** *Leo Lahti [aut], Sudarshan Shetty [aut], Felix GM Ernst [aut, cre]*
***Version:*** *0.98.9*
***Modified:*** *2021-04-10*
***Compiled:*** *2021-07-29*
***Environment:*** *R version 4.1.0 (2021-05-18), Bioconductor 3.14*
***License:*** *CC BY-NC-SA 3.0 US*
***Copyright:***
***Source:*** *https://github.com/microbiome/OMA*

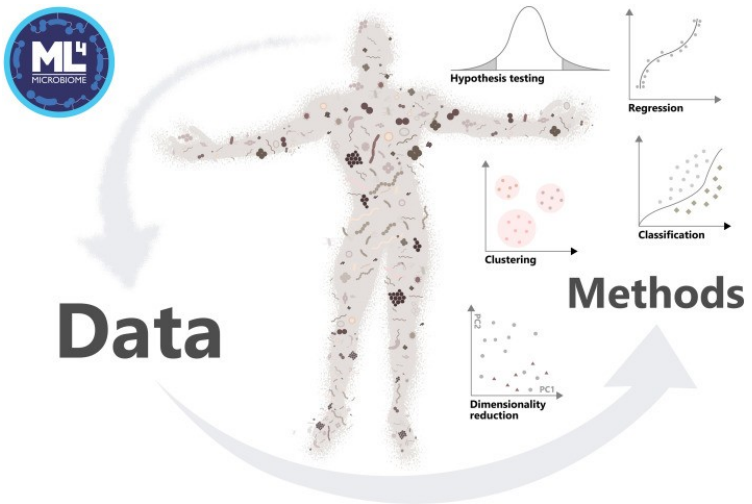



Figure source: Moreno-Indias et al. (2021) Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. Frontiers in Microbiology 12:11.

# Online tutorial (beta)

microbiome.github.io

Forthcoming courses & events:

June 20-23, Oulu, Finland

July 11-15, Nijmegen, The Netherlands

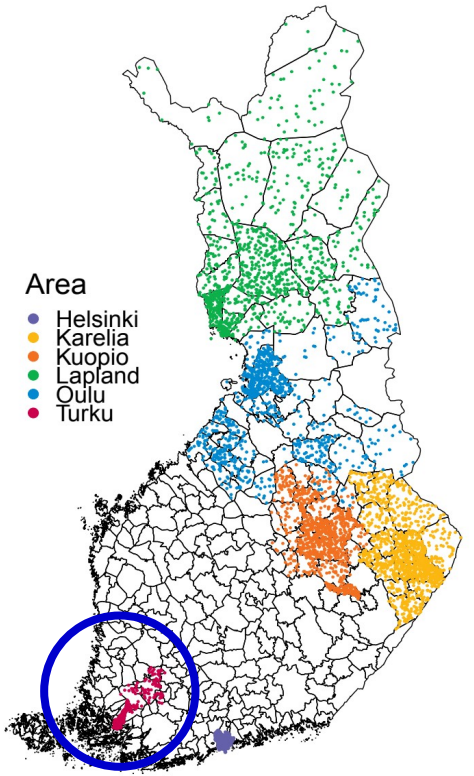
August 8-12, Pune, India (online / hybrid?)

Sep 26-30, CSC, Finland (online)

Oct 5-7, Barcelona, Spain (ML4microbiome)

Nov, Turku, Finland?


TURUN YLIOPISTO


Area
Helsinki
Karelia
Kuopio
Lapland
Oulu
Turku