Open data science in microbiome research

Grand challenges of data-Intensive science in microbiome & metagenome data analysis and training, October 14, 2021

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.

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

ML⁴ | MICROBIOME

@antagomir

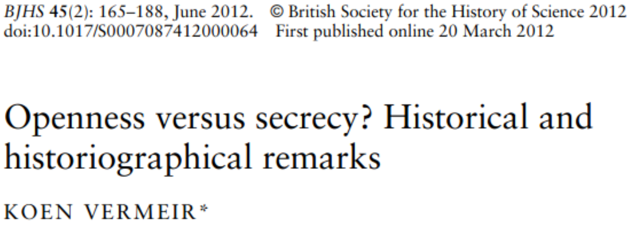
The demise of alchemy provides further evidence, if further evidence were needed, that what marks out modern science is not the conduct of experiments (alchemists conducted plenty of experiments), but the formation of a *critical community capable of assessing discoveries and replicating results*. Alchemy, 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

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.

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

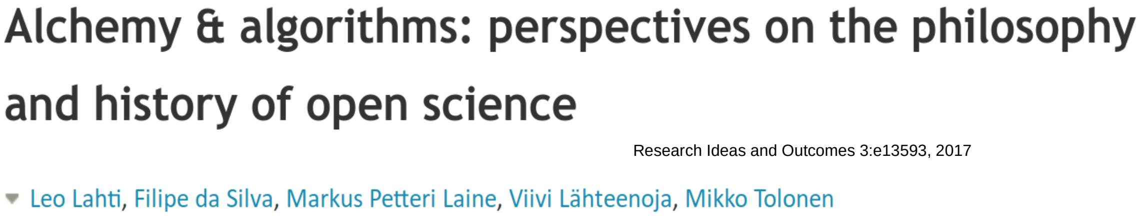
Source: Wikimedia Commons / Public domain



BJHS **45**(2): 165–188, June 2012. © British Society for the History of Science 2012
doi:10.1017/S0007087412000064 First published online 20 March 2012

Openness versus secrecy? Historical and historiographical remarks

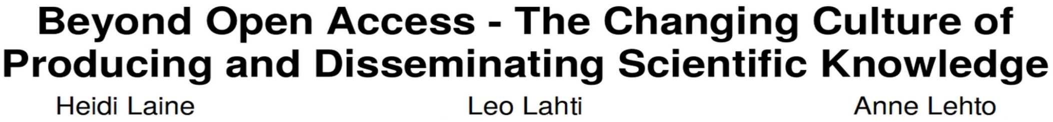
KOEN VERMEIR*



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

Research Ideas and Outcomes 3:e13593, 2017

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



Beyond Open Access - The Changing Culture of Producing and Disseminating Scientific Knowledge

Heidi Laine Leo Lahti Anne Lehto

Proceedings of the 19th International Academic Mindtrek Conference 2015. ACM New York, NY, USA. ISBN: 978-1-4503-3948-3 workflows

community

reproducibility

ML4 | MICROBIOME

ml4microbiome.eu

(some) elements of microbiome data science

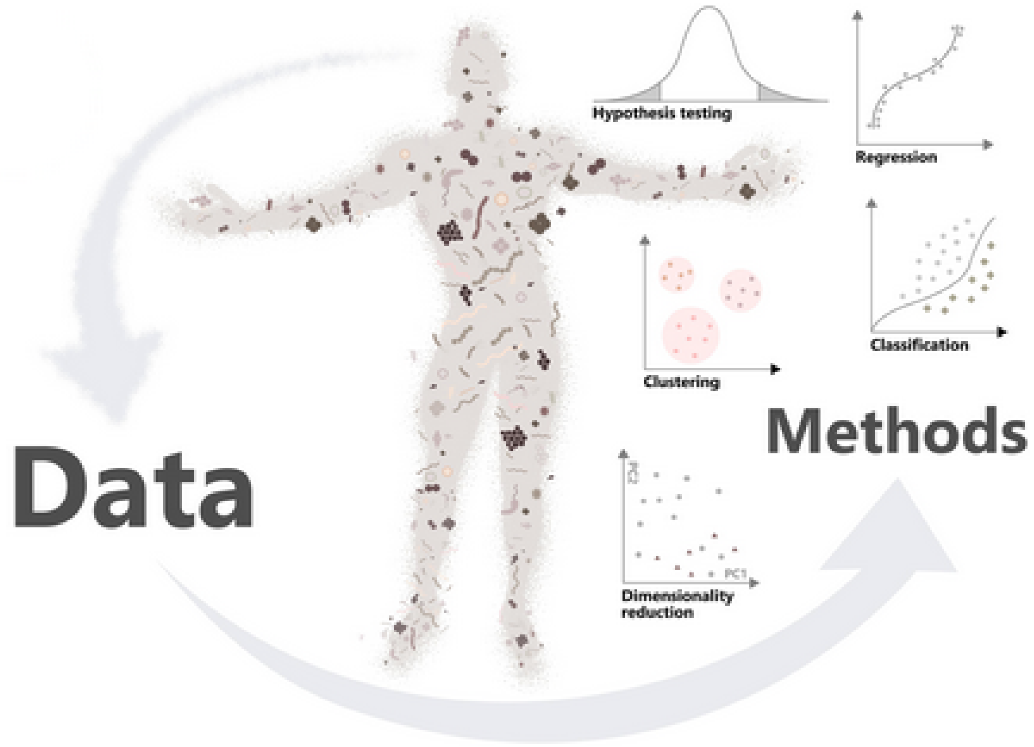
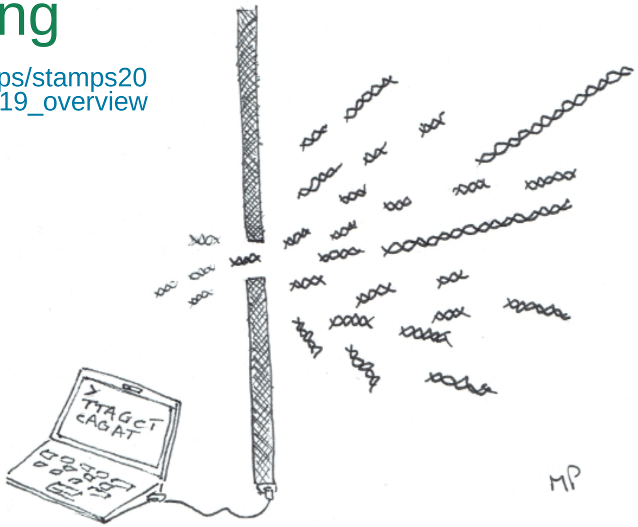
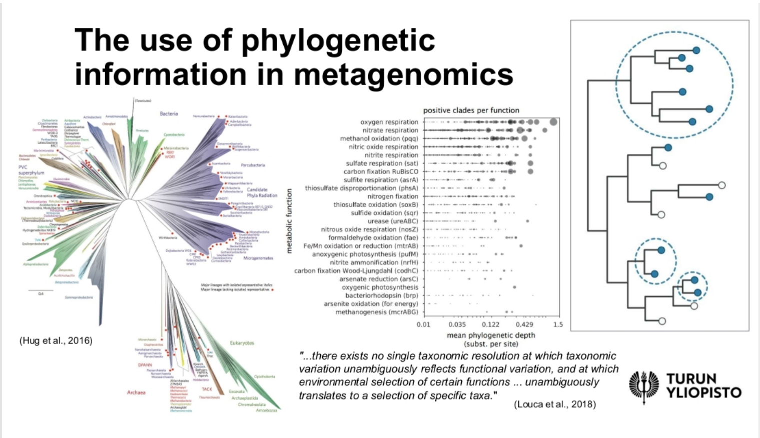


Figure source: Moreno-Indias *et al.* (2021) Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. Frontiers in Microbiology 12:11.

Finite sampling

https://github.com/mblstamps/stamps2019/blob/master/STAMPS2019_overview_Pop.pdf





The use of phylogenetic information in metagenomics

positive clades per function
metabolic function
oxygen respiration
nitrate respiration
methanol oxidation (pqq)
nitric oxide respiration
nitrite respiration
sulfate respiration (sat)
carbon fixation RuBisCO
sulfite respiration (asrA)
thiosulfate disproportionation (phsA)
nitrogen fixation
thiosulfate oxidation (soxB)
sulfide oxidation (sqr)
urease (ureABC)
nitrous oxide respiration (nosZ)
formaldehyde oxidation (fae)
Fe/Mn oxidation or reduction (mtrAB)
anoxygenic photosynthesis (pufM)
nitrite ammonification (nrfH)
carbon fixation Wood-Ljungdahl (codhC)
arsenate reduction (arsC)
oxygenic photosynthesis
bacteriorhodopsin (brp)
arsenite oxidation (for energy)
methanogenesis (mcrABG)
0.01
0.035
0.122
0.429
1.5
mean phylogenetic depth
(subst. per site)
Bacteria
Candidate Phyla Radiation
Microgenomates
Eukaryotes
Archaea
Opisthokonta
Amoebozoa

(Hug et al., 2016)

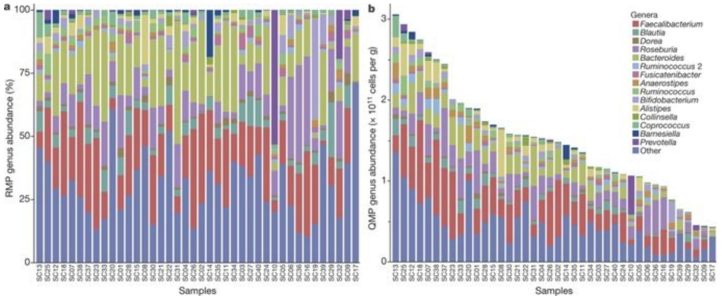
"...there exists no single taxonomic resolution at which taxonomic variation unambiguously reflects functional variation, and at which environmental selection of certain functions ... unambiguously translates to a selection of specific taxa."

(Louca et al., 2018)

TURUN YLIOPISTO

Compositionality

Vandeputte et al. Nature 551:507-511, 2017



The variety of study types

Data preprocessing

Case-control studies

Interventions

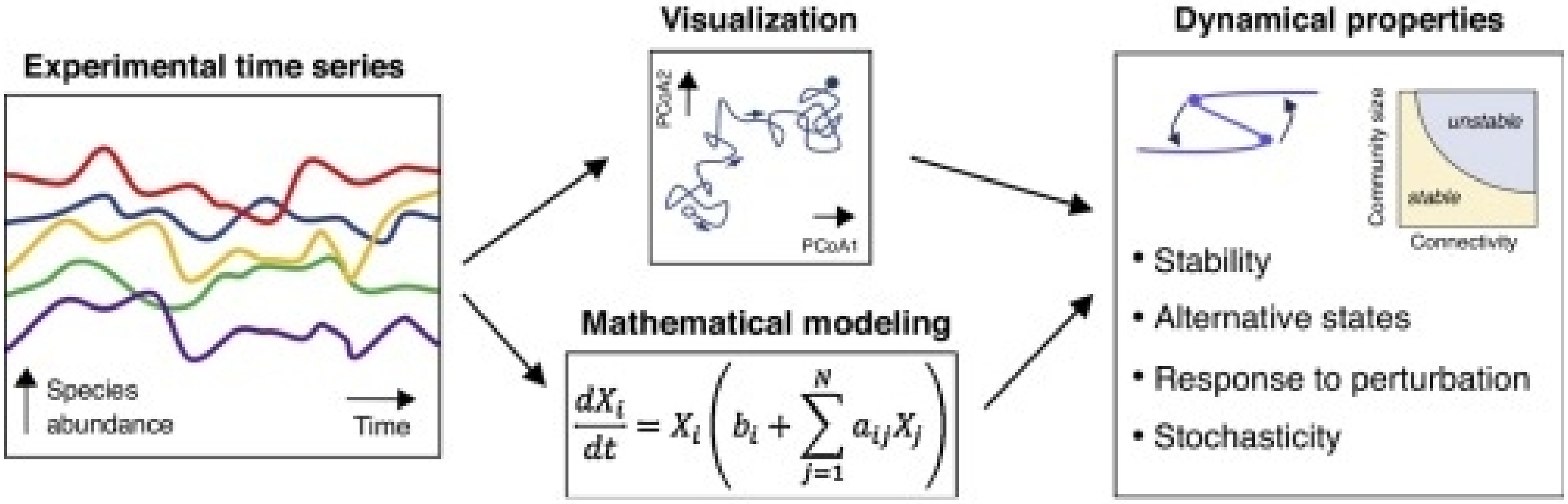
Cross-sectional analysis

Prospective analysis

Longitudinal dynamics

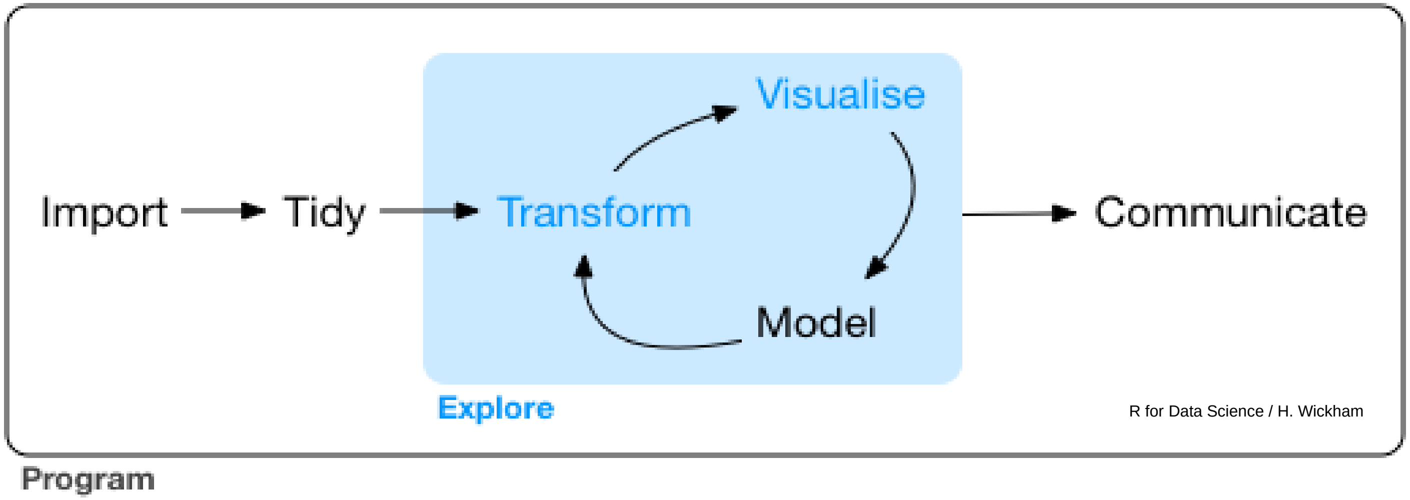
Multi-omics

Microbiome research is *data-intensive* and relies on a heterogeneous array of *sophisticated computational techniques*



Gonze et al. Curr. Op. Microbiol. 2018. Microbial communities as dynamical systems.

Reproducible workflows improve transparency and robustness



R for Data Science / H. Wickham

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

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

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

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

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

Overview of generic* metagenomics workflow

*This is generic; specific workflows can vary on the order of steps here and how they are done.

When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.

sequencing facility → fastq files → demultiplex → quality filter/trim → fasta files

fastq files

```
@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...
```

might be done by sequencing facility

demultiplex
(split samples by barodes)

Some tools:
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

fastqc/multiqc

quality filter/trim
(remove adapters/**primers**)

Some tools:
- trimmomatic
- bbduk.sh (bbtools suite of tools)

fasta files

```
>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
```

no-assembly path

read-based analysis

Some tools:
- TIPP/SEPP
- metaphlan2
- humann2
- sourmash
- kraken

Count Table

	Sample_A	Sample_B	. . .
obj_1	0	428	. . .
obj_2	306	323	. . .
obj_3	217	1	. . .
. . .	. . .	. . .	. . .

Analysis

Some tools:
- phyloseq
- Breakaway
- DivNet
- CORNCOB
- SpiecEasi
- MaAsLin
- DESeq2

assembly path

consider testing assemblies with and w/o

digital normalization

Some tools:
- bbnorm
- diginorm

MetaQUAST is a great tool for comparing assemblies

(co)-assembly

Some assemblers and tools:
- Megahit (assembler)
- SPAdes (assembler)
- idba-ud (assembler)
- MetAMOS (assembler and analysis pipeline)
- MetaCompass (reference-guided)
- MetagenomeScope (visualize assembly graphs)

map individual sample reads to (co)-assembly

Generate coverage information (mapping)

Some tools:
- bowtie2
- bwa

Recovering genomes from metagenomes

→ **A note on MAGs:**
MAGs (metagenome-assembled genomes) are not the same thing as isolate genomes. They are composite representative genomes of closely related genomic lineages.

Some tools:
- anvi'o (interactive manual curation of bins; and much more)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- COCACOLA (kmer-based, coverage-based, and incorporates paired-read linkage of contigs)
- MetaBAT2 (kmer-based and coverage-based binning tool)
- BinSanity (primarily coverage-based, optional second round kmer-based binning tool)
- checkm (genome-level taxonomy; and much more)
- DASTool (a tool for evaluating bins recovered by different methods)
- DESMAN (tool aimed at resolving strains)

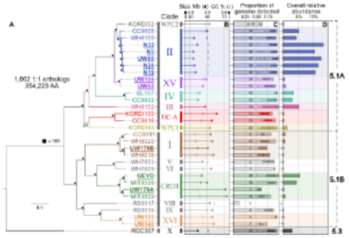
Gene calling Functional/taxonomic profiling

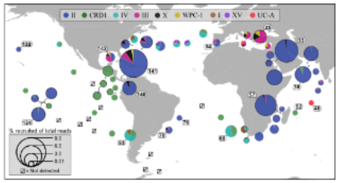
Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

Some common genomics stuff

**Phylogenomics
Comparative genomics
Pangenomics
Env. distributions**

Some tools:
- anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL)
- PanOCT (identifies orthologs utilizing synteny information)
- StrainPhlAn/PanPhlAn (tools for strain-level analyses)
- MUSCLE (alignment software)
- FastTree (very fast, pseudo-maximum likelihood tree builder)
- RAxML (maximum likelihood tree builder)
- Mauve (whole-genome alignment)





astrobiomike.github.io

Happy Belly Bioinformatics

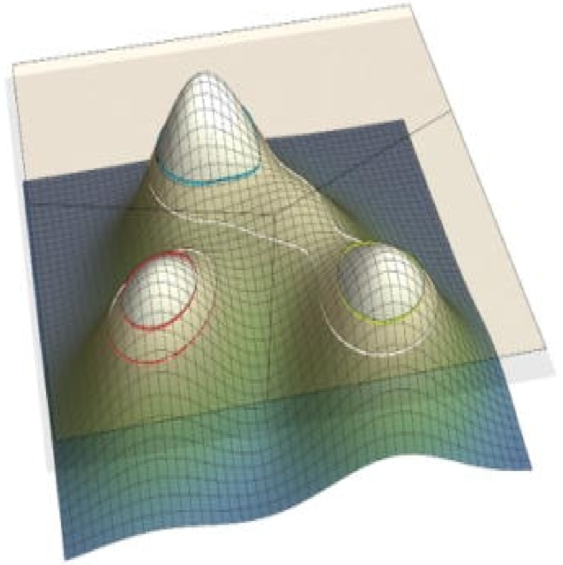
JOSE 10.21105/jose.00053

AstrobioMike

Orcid: 0000-0001-7750-9145

Lee, (2019). Happy Belly Bioinformatics: an open-source resource dedicated to helping biologists utilize bioinformatics. Journal of Open Source Education, 4(41), 53, https://doi.org/10.21105/jose.00053

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



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

$$2 \text{ x } 6^6 = 93312$$

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

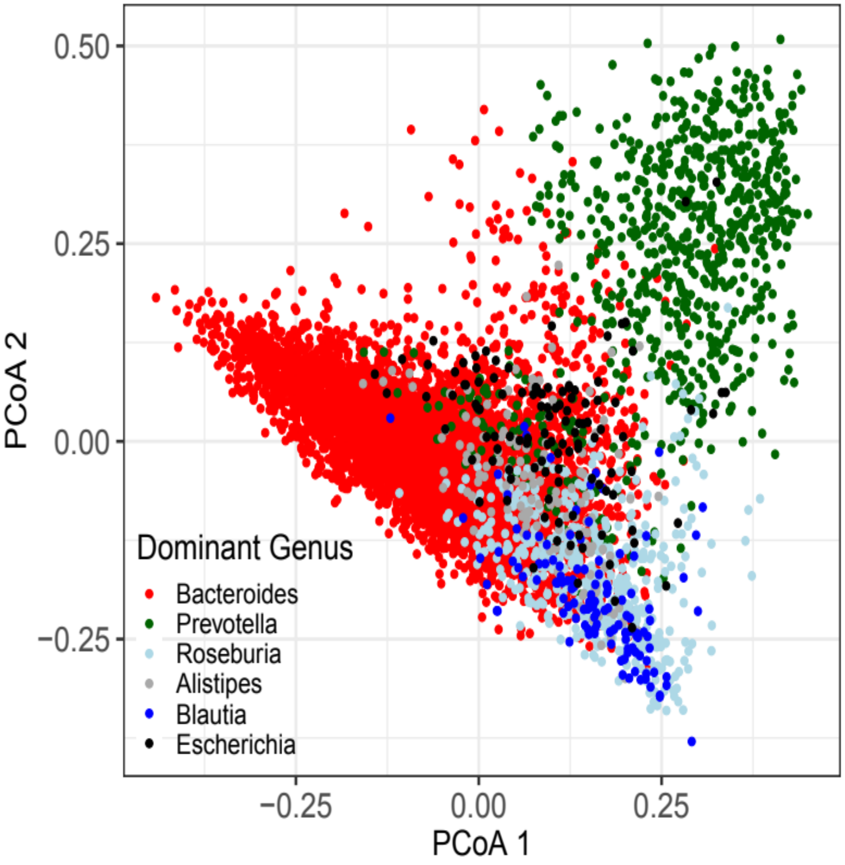
Clustering method
- Hierarchical / Ward
- Hierarchical / Complete
- Gaussian mixture
- DMM
- PAMR
- K-means

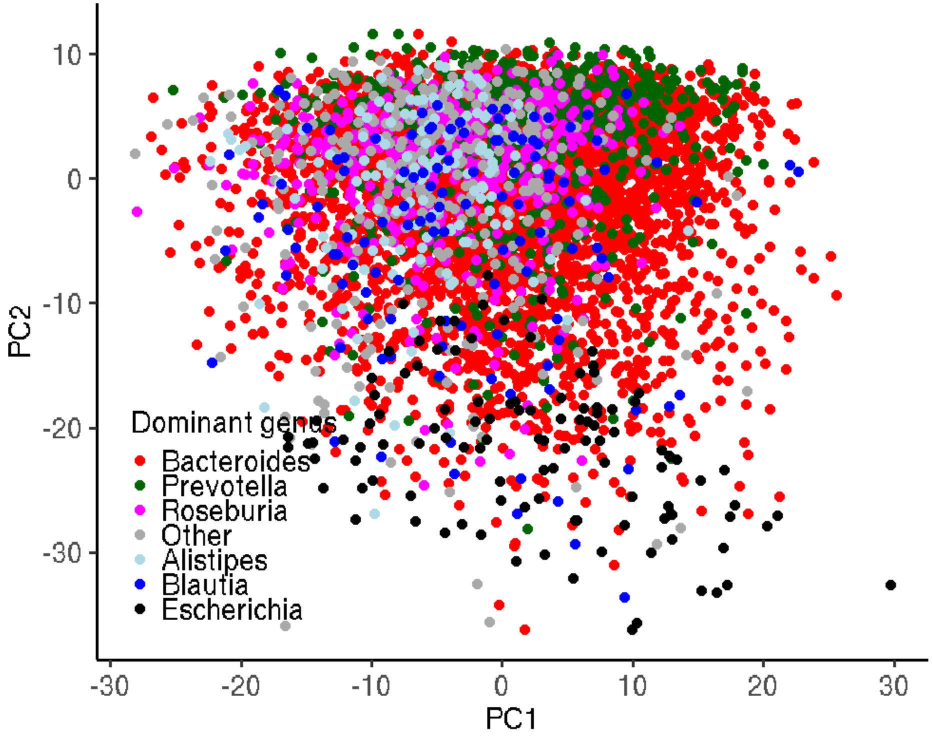
Regulation
- Calinski-Harabasz
- Dirichlet Process
- Silhouette Index
- AIC
- BIC
- DIC

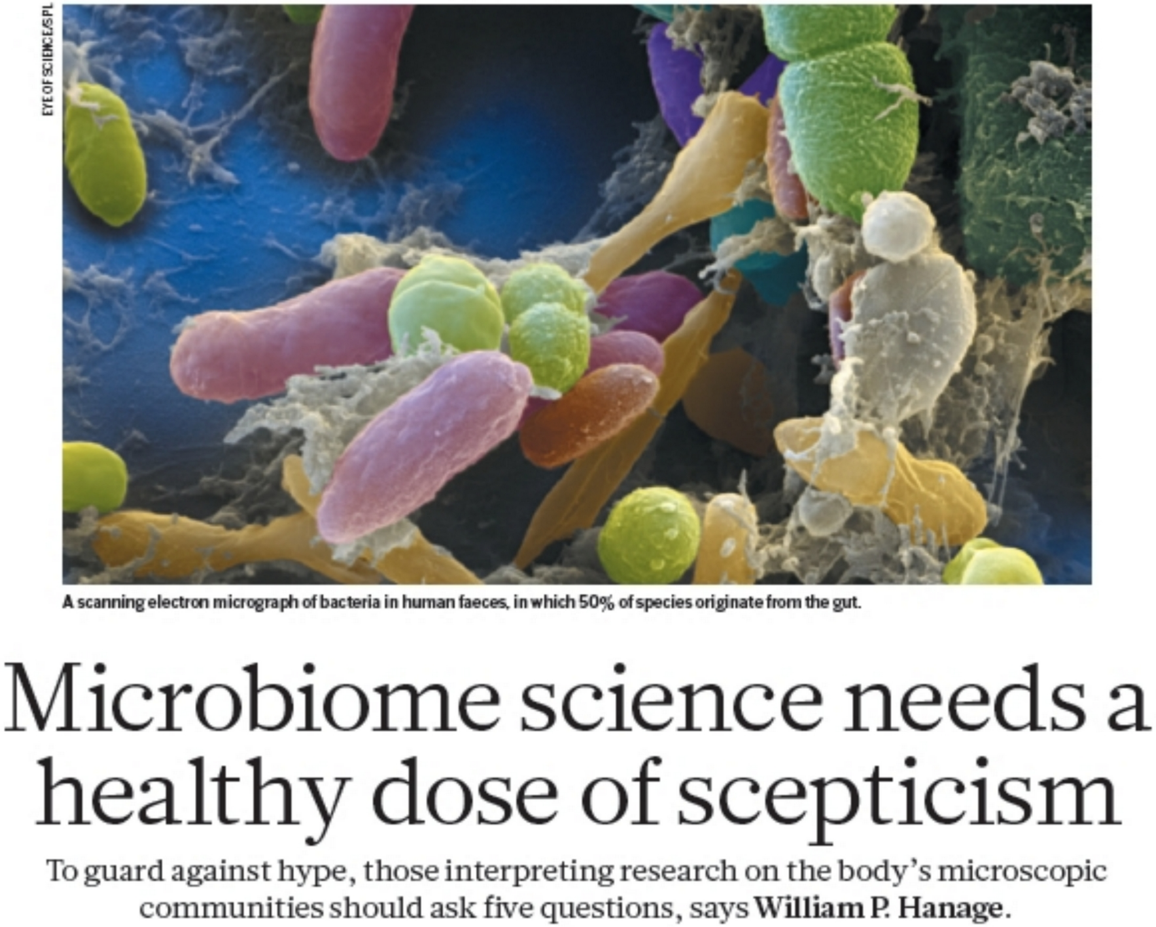
Walk-through example in R/Bioc by Holmes & McMurdie
http://statweb.stanford.edu/~susan/papers/EnterotypeRR.html

PCoA or PCA –
different methods, different results?

FINRISK – Salosensaari et al. Nat. Comm. 2021







EYE OF SCIENCE/SPL

A scanning electron micrograph of bacteria in human faeces, in which 50% of species originate from the gut.

Microbiome science needs a healthy dose of scepticism

To guard against hype, those interpreting research on the body's microscopic communities should ask five questions, says **William P. Hanage**.

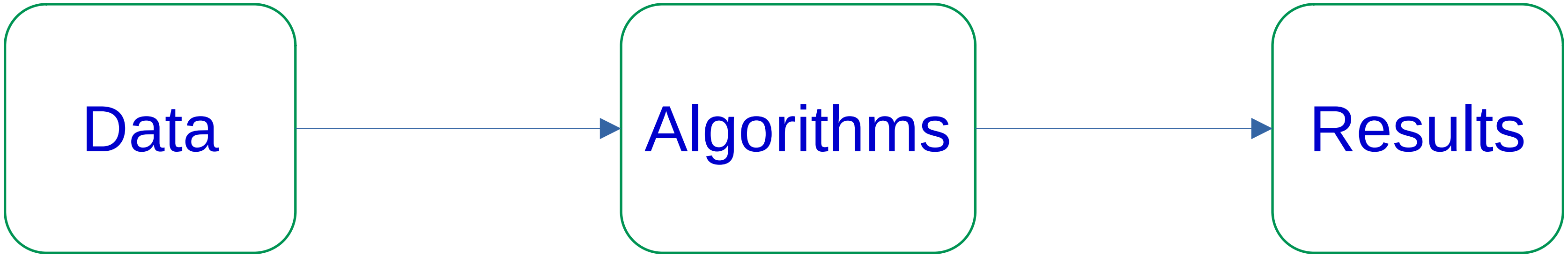
Comment August 2014 Nature

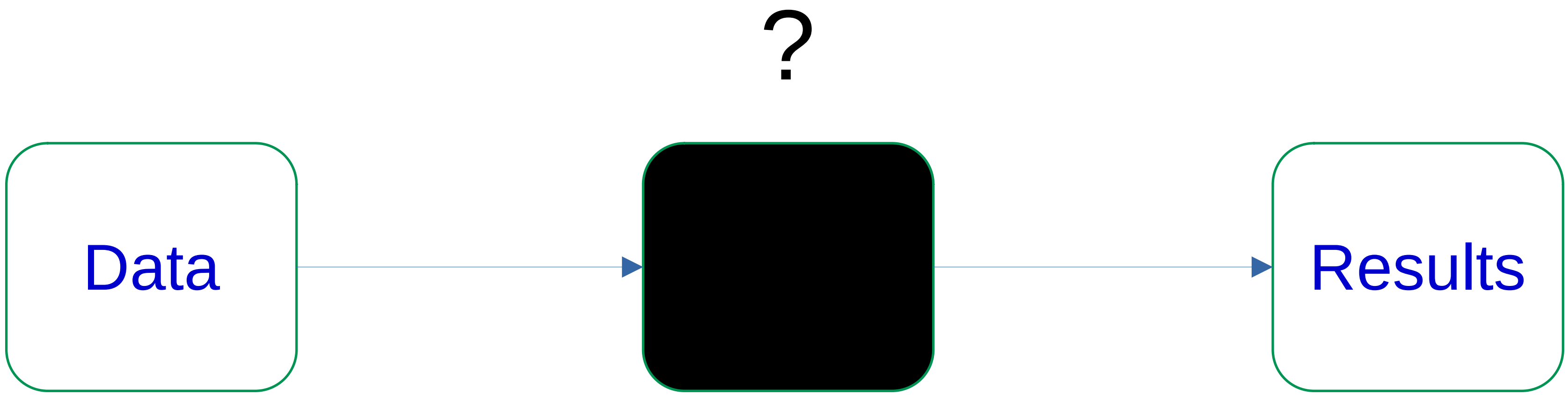
The influence of hidden researcher decisions in applied microeconomics

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.


Data
Algorithms
Results



?

Data

Results

RESEARCH PRIORITIES

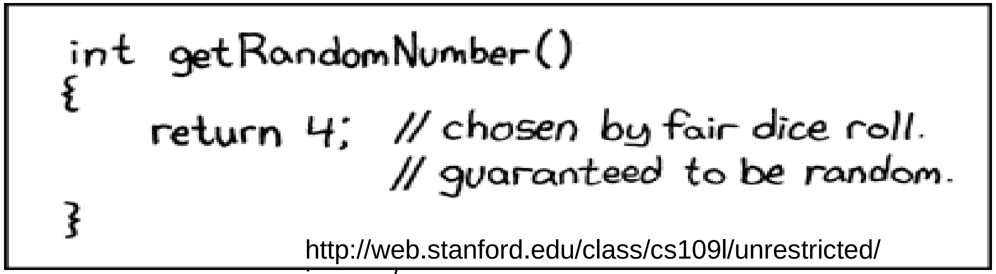
Shining Light into Black Boxes

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

How to Make More Published Research True

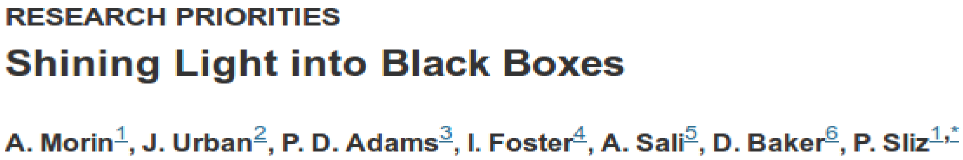
John P. A. Ioannidis

Published: October 21, 2014 • DOI: 10.1371/journal.pmed.1001747



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

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



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,*]

“*I have begun to think that no one ought to publish biometric results, without lodging a well arranged and well bound manuscript copy of all his data, in some place whereit should be accessible, under reasonable restrictions, to those who desire to verify his work.*”

Francis Galton (1901), *Biometrika* 1:1, pp. 7-10.

Data silo

open data science ecosystems

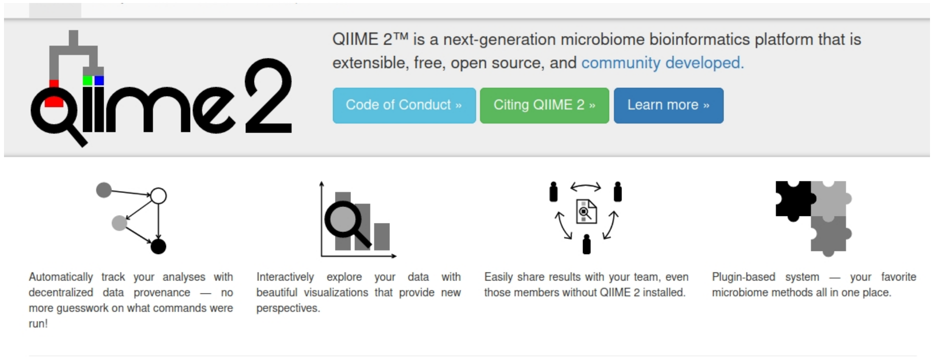
mothur

Welcome to the website for the mothur project, initiated by Dr. Patrick Schloss and his research group in 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. In February 2009 we released the first 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, PacBio, IonTorrent, 454, and Illumina (MiSeq/HiSeq). 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

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

This site is maintained by Pat Schloss
© 2008-2019



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

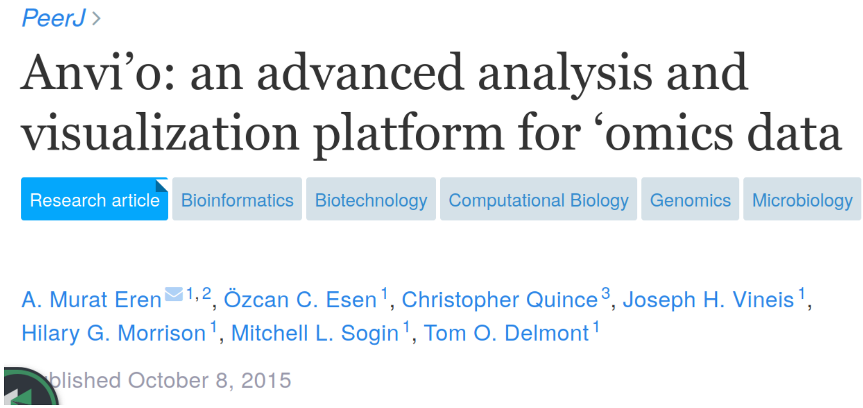
Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!

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.

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

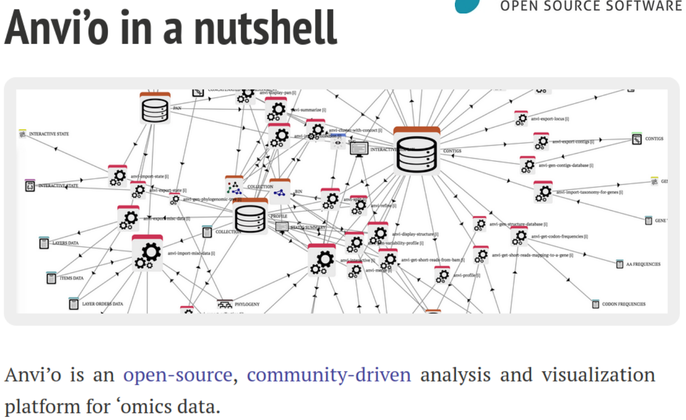


PeerJ ›

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

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]

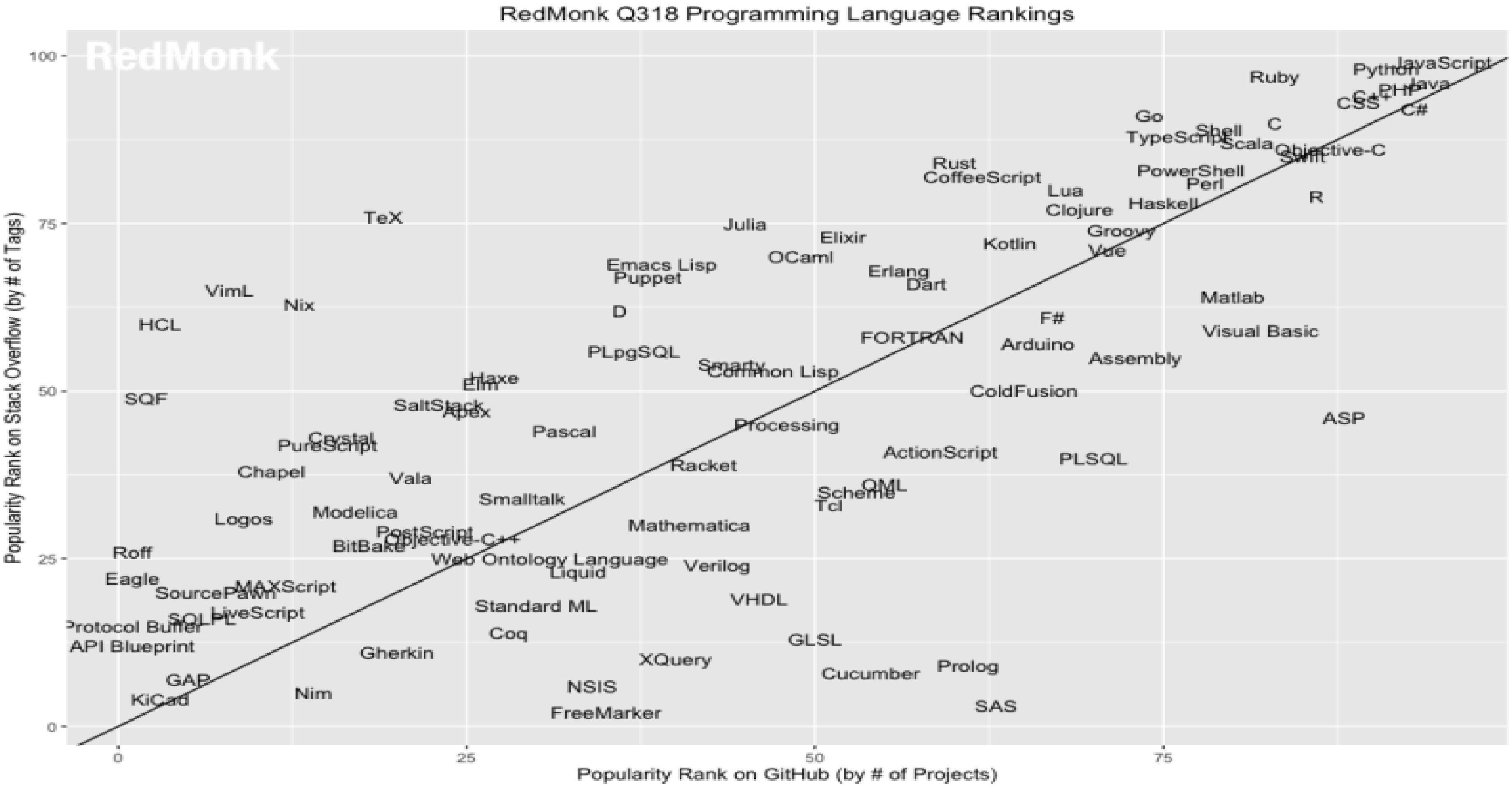
blished October 8, 2015



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


RedMonk Q318 Programming Language Rankings

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

What is Bioconductor ?

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

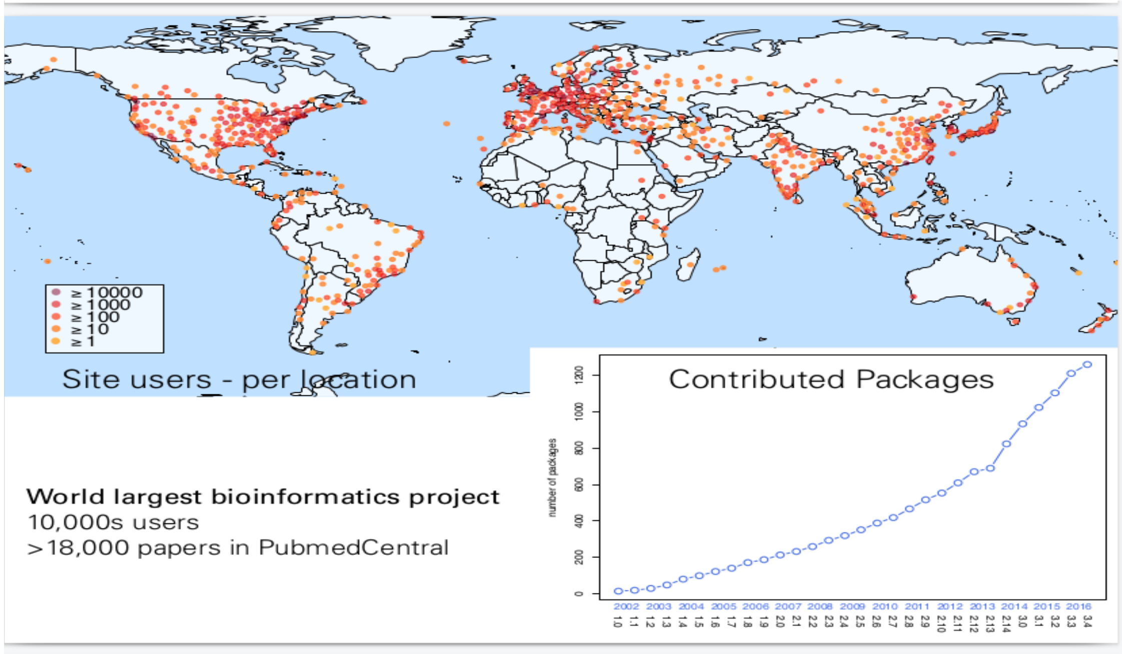
Started 2001 as a platform for analysis & understanding of microarray data

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

What is Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS ?

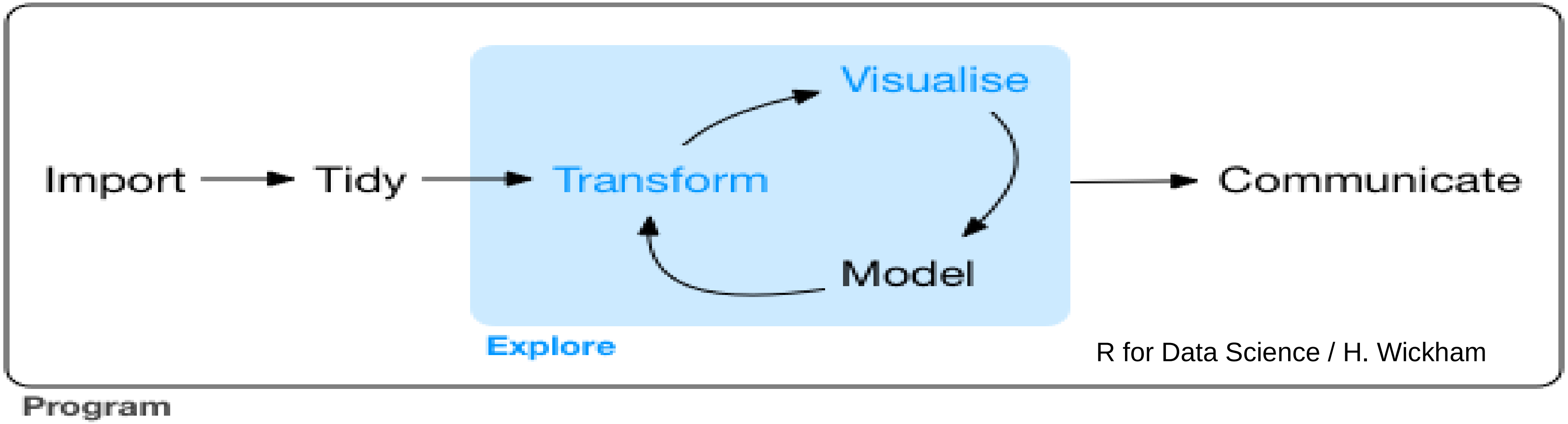
Principally a collaborative software development project

But it is also:

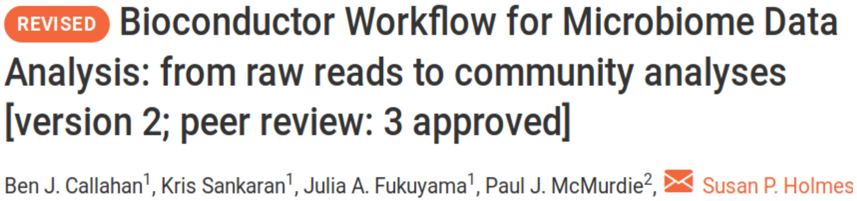
- a software repository
- a bioinformatics support site
- data repository
- publisher for supplementary materials
- source for tutorials and instructional documentation

Managed and maintained by a core team of ~6 people, with contributions coming from all over the world

Data science workflow

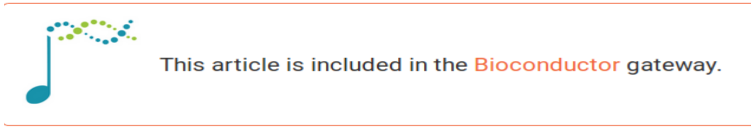


R for Data Science / H. Wickham

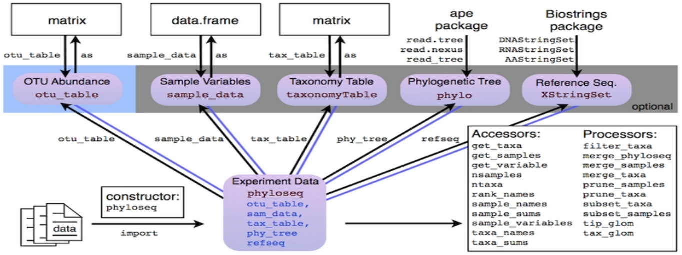


REVISED Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses [version 2; peer review: 3 approved]

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

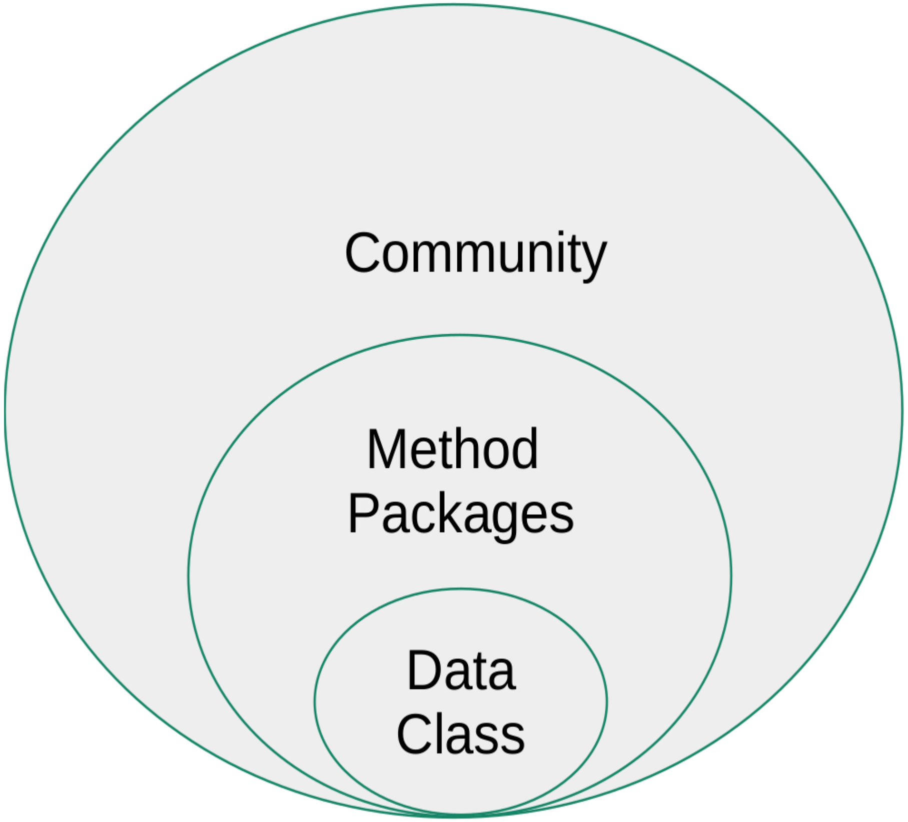


This article is included in the Bioconductor gateway.

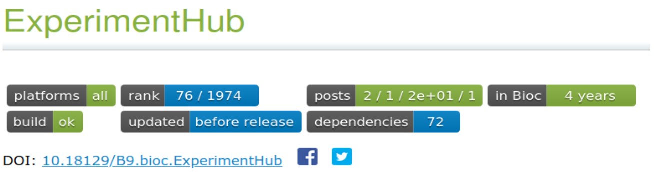


popular data containers
support collaborative research
and methods development

Reduce overlapping efforts, improve interoperability, ensure sustainability.



Data packages

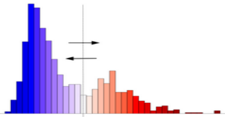

ExperimentHub

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

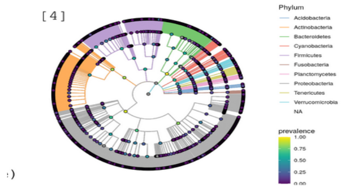
DOI: 10.18129/B9.bioc.ExperimentHub

mia – microbiome analysis

getDiversity(x)
calculateDMM(x)



miaViz - Visualization

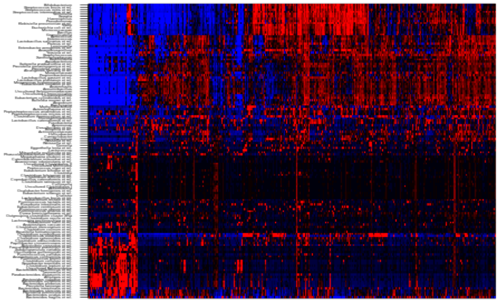


```
[4] miaViz::plotRowTree(tse)
```

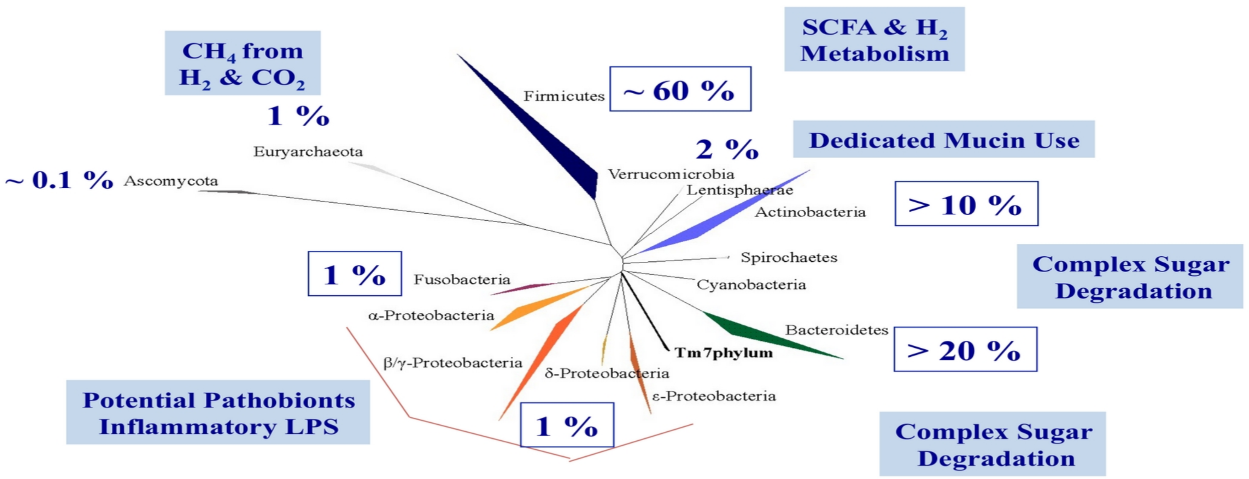
Package ecosystem

Special properties of microbiome data

- Sparse
- Compositional
- Non-Gaussian
- Overdispersed
- Discrete
- Complex
- Stochastic
- Multi-level



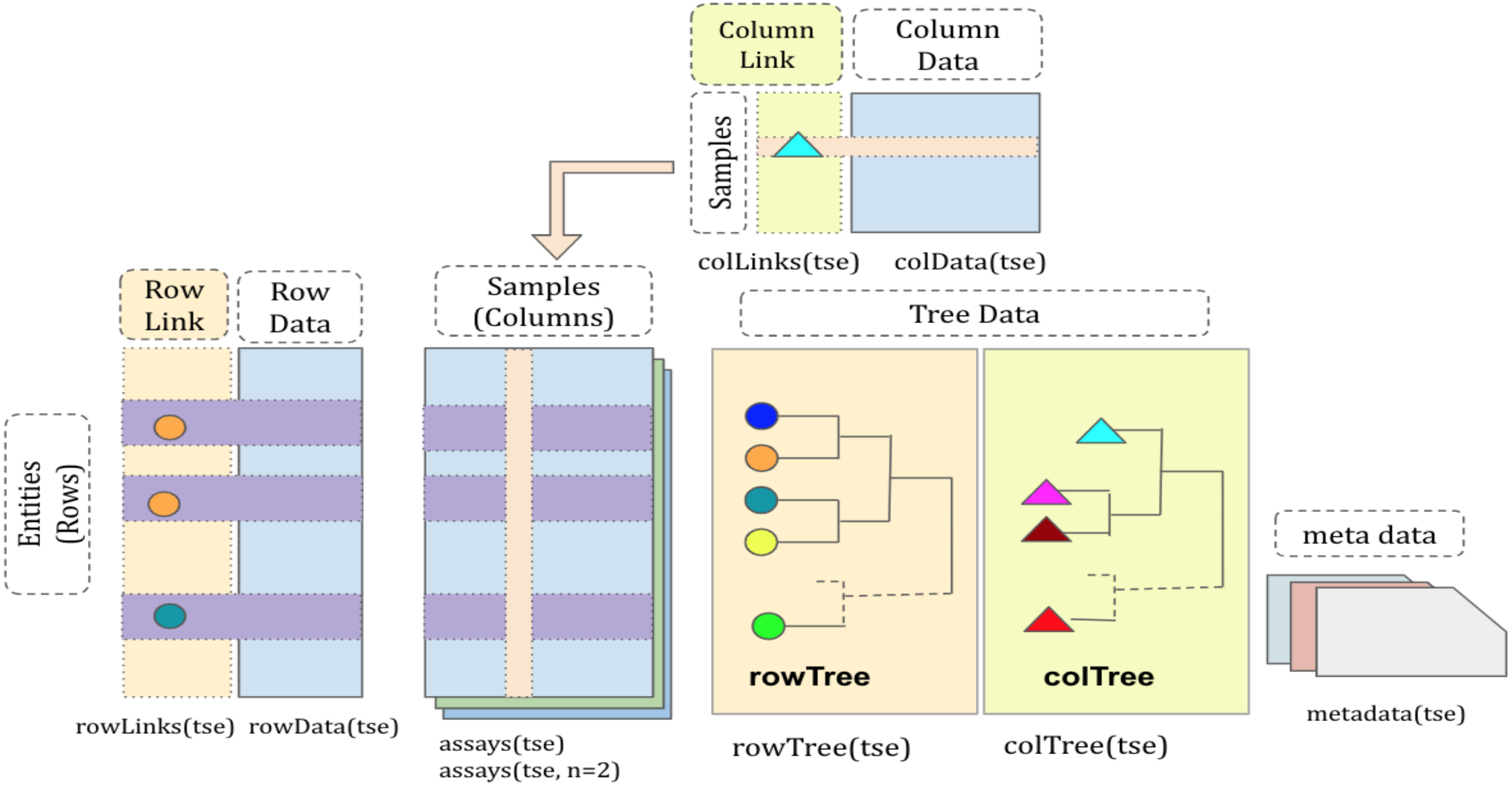
Colonic Climax Communi



Zoetendal EG, EE Vaughan & WM de Vos (2006) Mol Microbiol 59: 1639

Lay C, L Rigottier-Gois, K Holmstrom, M Rajilic, EE Vaughan, WM de Vos, MD Collins, R Their, P Namsolleck, M Blaut & J Dore (2005) AEM 71: 4153

Anatomy of TreeSummarizedExperiment



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]

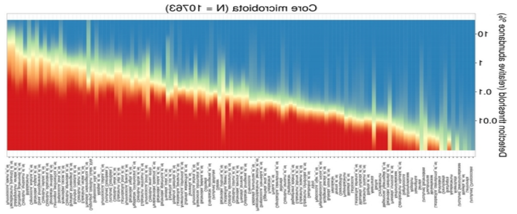
Property	Algorithms	Consequence
Analytical solution	Concat. PCA, CCA, CoIA, MFA, PTA, Statico/Costatis	Methods with analytical solutions generally run much faster than those that require iterative updates, optimization, or Monte Carlo sampling. They tend to be restricted to more classical settings, however.
Require covariance estimate	Concat. PCA, CCA, CoIA, 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, CoIA, 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.

microbiome R package

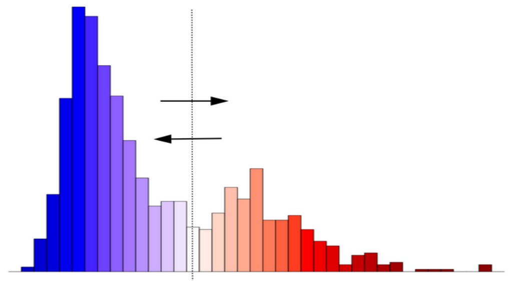
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

chat on gitter | build passing | codecov 24% | PRs welcome

Core & prevalence
prevalence(x)
core(x)
core_members(x)

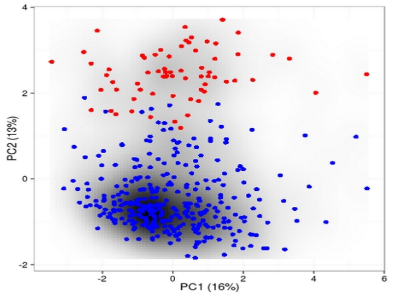


Stability & resilience



Alpha & beta diversity

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



Transformations

transform(x, “compositional”)
transform(x, “clr”)
transform(x, “log10p”)
transform(x, “hellinger”)
transform(x, “identity”)

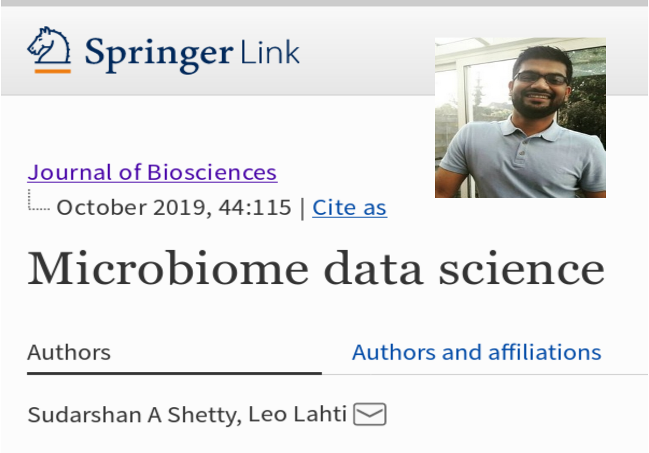
Community
- Online tutorials
- Mailing list
- Gitter chat
- Example data
- Workshops

Quality control
- continous integration
- unit tests

microbiome.github.io

A survey for 16S
Github.com/microsud/
Tools-Microbiome-Analysis

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



Journal of Biosciences
October 2019, 44:115 | Cite as

Microbiome data science

Authors | Authors and affiliations

Sudarshan A Shetty, Leo Lahti

HUMAN GUT MICROBIOME ATLAS

Open access resource for human microbiome. SciLifeLab, King´s College London, INRAE

Workflow approach:
knitting together open data, methods & application

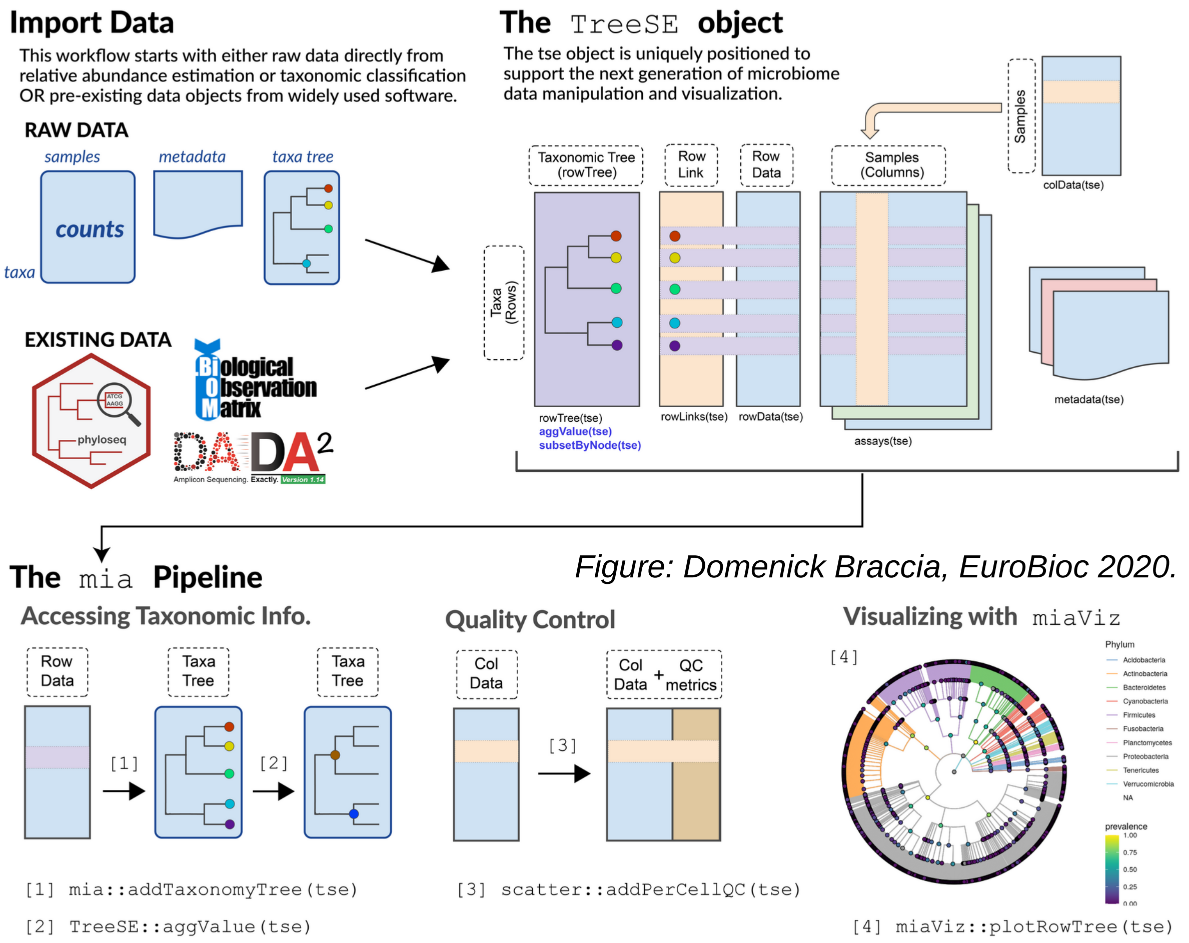


Figure: Domenick Braccia, EuroBioc 2020.

Check the poster
F1000 / EuroBioC!

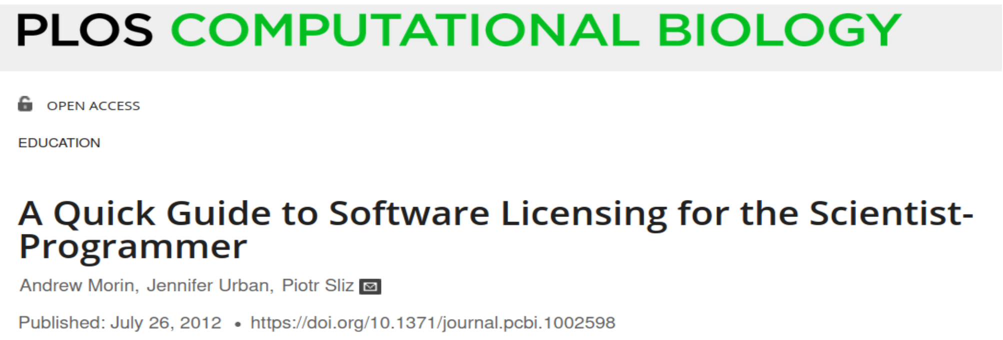
- comprehensive
- extendable
- reproducible
- collaborative
- transparent

elements of computational workflows

(transparent?) data

(open) algorithms

(reproducible) reporting



PLOS COMPUTATIONAL BIOLOGY

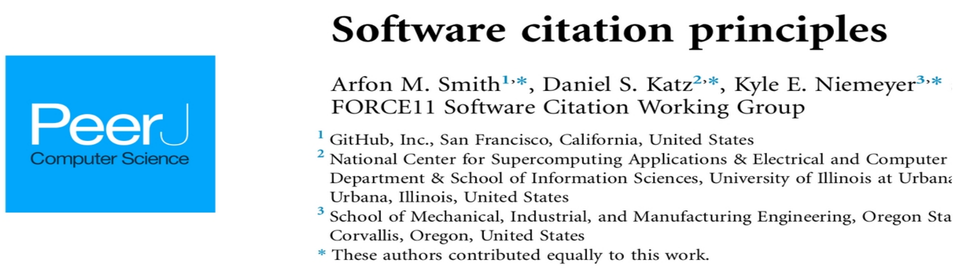
OPEN ACCESS

EDUCATION

A Quick Guide to Software Licensing for the Scientist-Programmer

Andrew Morin, Jennifer Urban, Piotr Sliz

Published: July 26, 2012 • https://doi.org/10.1371/journal.pcbi.1002598

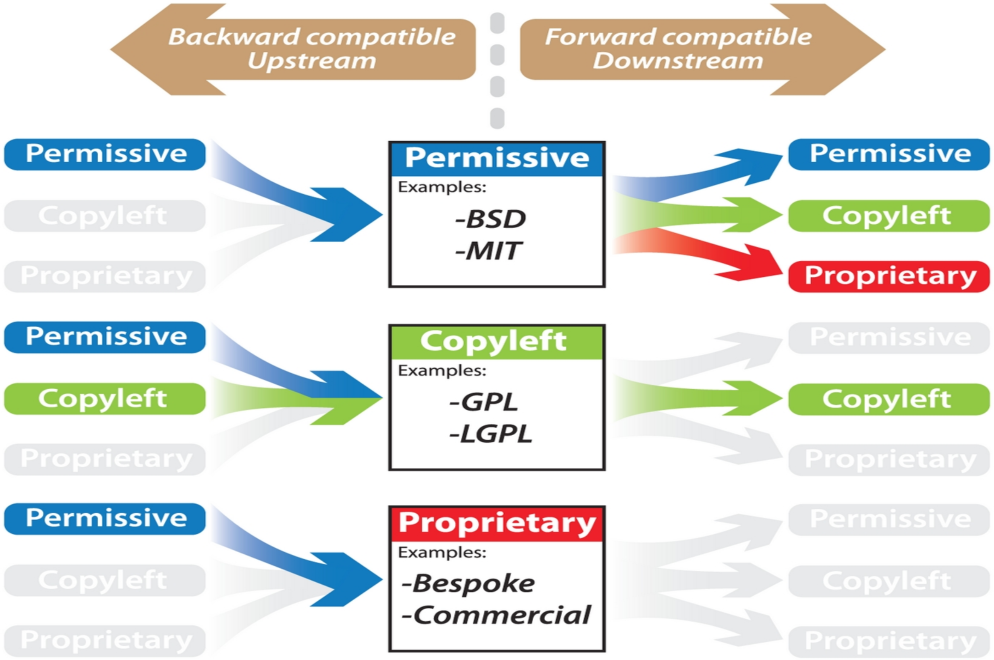


PeerJ Computer Science

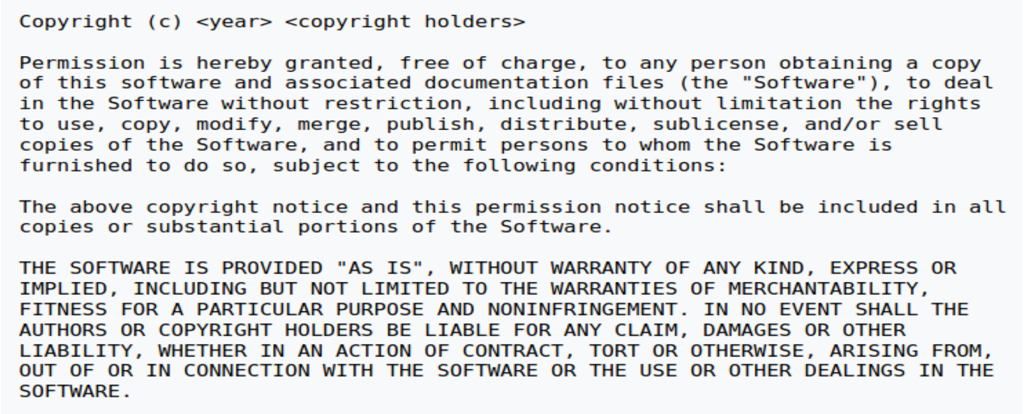
Software citation principles

Arfon M. Smith[1,*], Daniel S. Katz[2,*], Kyle E. Niemeyer[3,*]
FORCE11 Software Citation Working Group

[1] GitHub, Inc., San Francisco, California, United States
[2] National Center for Supercomputing Applications & Electrical and Computer Department & School of Information Sciences, University of Illinois at Urbana Urbana, Illinois, United States
[3] School of Mechanical, Industrial, and Manufacturing Engineering, Oregon Sta Corvallis, Oregon, United States
[*] These authors contributed equally to this work.



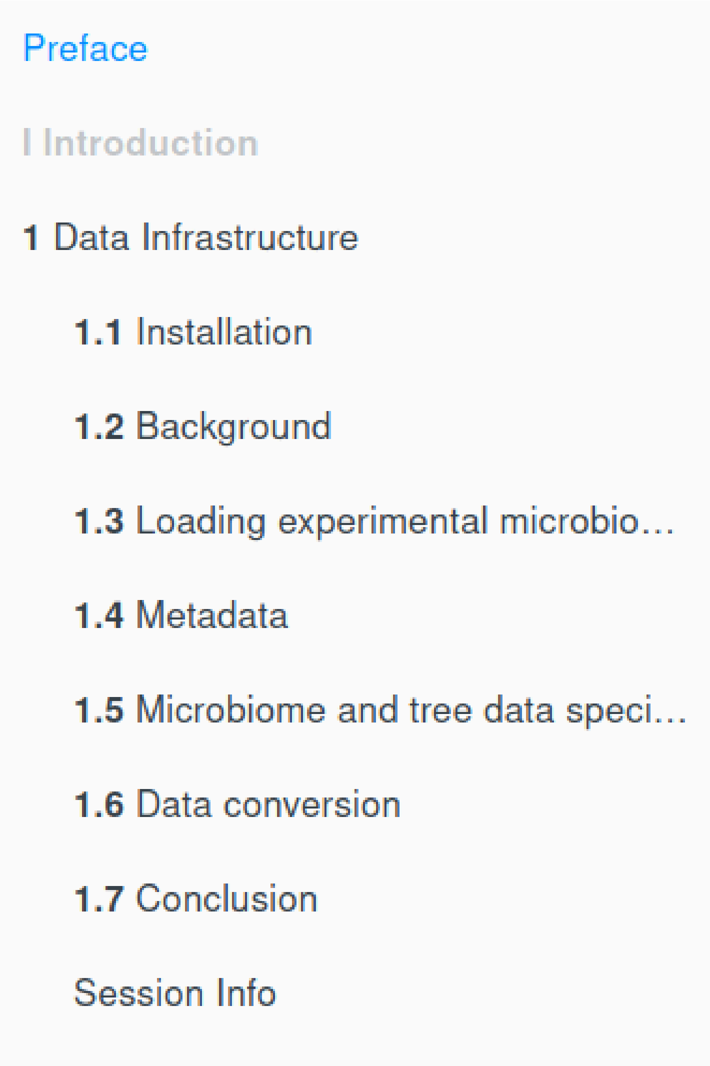
MIT License

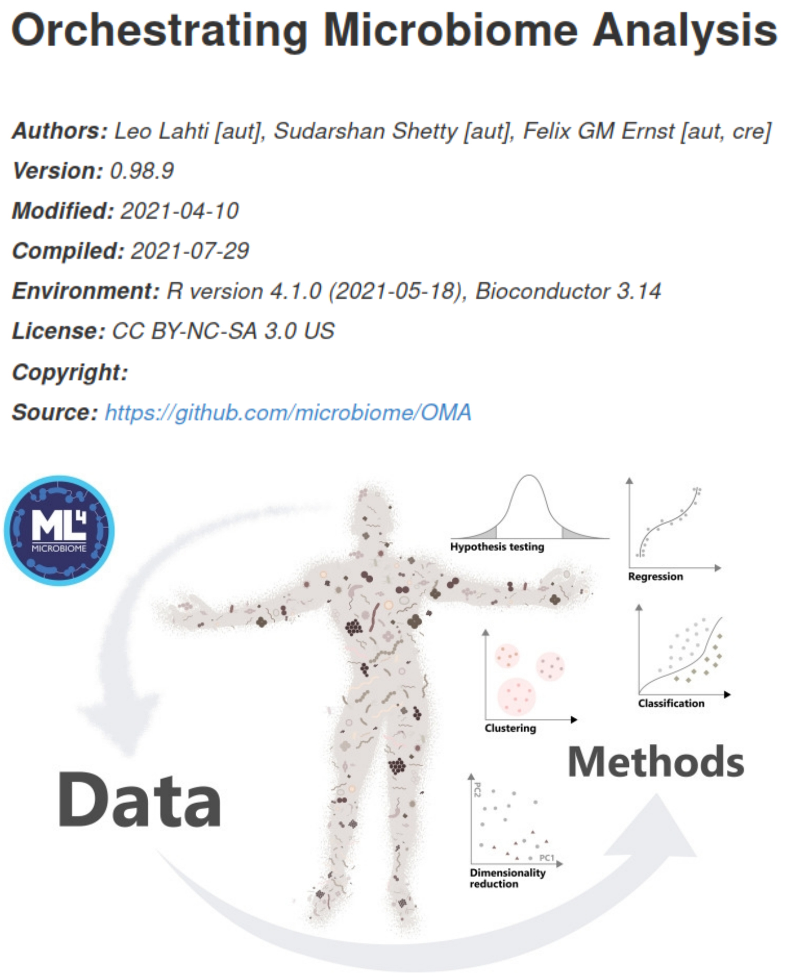


```
Copyright (c) <year> <copyright holders>

Permission is hereby granted, free of charge, to any person obtaining a copy
of this software and associated documentation files (the "Software"), to deal
in the Software without restriction, including without limitation the rights
to use, copy, modify, merge, publish, distribute, sublicense, and/or sell
copies of the Software, and to permit persons to whom the Software is
furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all
copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
SOFTWARE.
```

Orchestrating Microbiome Analysis with R/Bioc

microbiome.github.io





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.

Training events in microbiome data science

ML4Microbiome Symposium: Grand Challenges of Data-Intensive Science in microbiome data analysis and training (Oct 14)

ML4Microbiome Training School, Sep/Oct 2021

Workshop on modeling microbial community time series. Leuven, **Belgium**, November, 2021

Brain, Bacteria and Behaviour: Understanding the Gut-Brain Axis online summer school, The **Netherlands**, July 2021

NORBIS Summer School; National research school in bioinformatics, biostatistics and systems biology, **Norway**, Aug 2021

Microbiome Data Analysis Workshop - Hasselt University, Limburg, **Belgium**, Apr 2021

Techniques for skin microbiome research - Savitribai Phule Pune University, Pune, **India**, Jan 2021

Modern statistics for microbiome bioinformatics - Pune, **India**, Dec 2019

Intestinal microbiome of humans and animals. Wageningen University and Research Center, The **Netherlands**, Oct 2019

Microbiome data science. **Singapore** Centre for Environmental and Life Science Engineering Sep, 2019

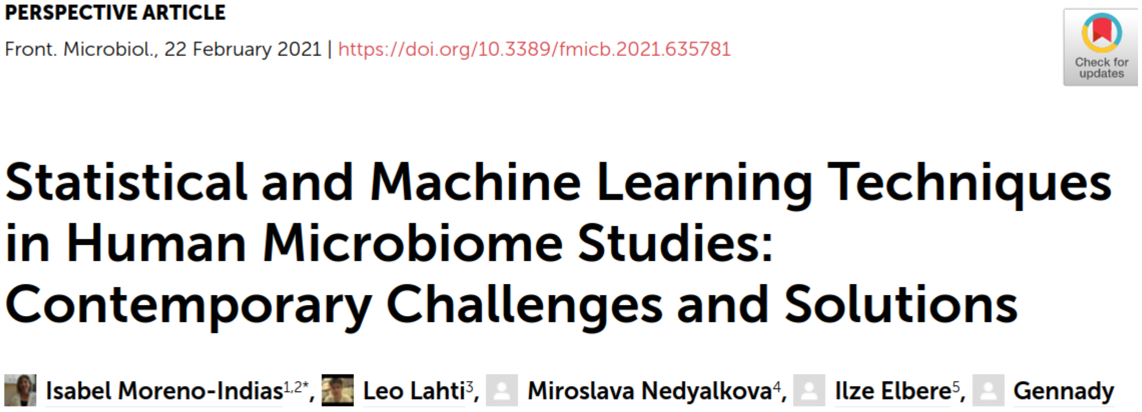
Statistical techniques in microbiome bioinformatics - Sep 2019 Radboud University Nijmegen, The **Netherlands**

International summer school on microbial community modeling - Sep 2019 KU Leuven, **Belgium**

International spring school on open microbiome data analysis - 2018 Wageningen, The **Netherlands**

International summer school on microbial network analysis - 2017 KU Leuven, **Belgium**

datascience.utu.fi | ml4microbiome.eu



PERSPECTIVE ARTICLE

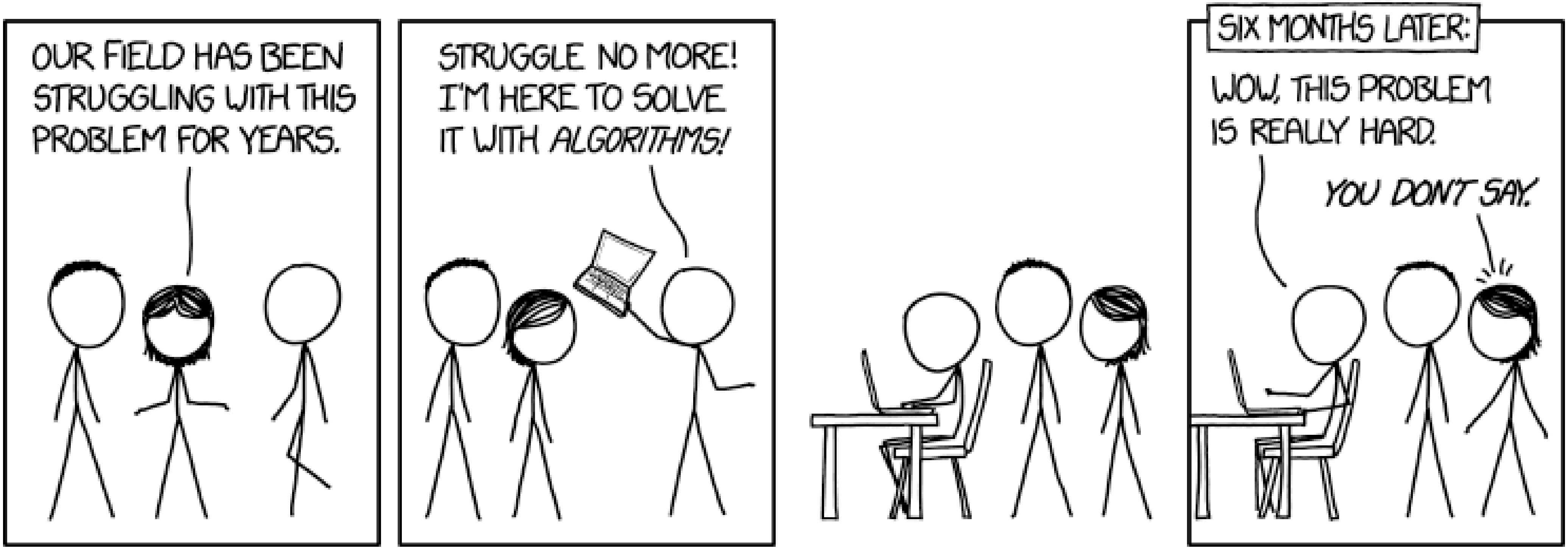
Front. Microbiol., 22 February 2021 | https://doi.org/10.3389/fmicb.2021.635781

Check for updates

Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions

Isabel Moreno-Indias[1,2*], Leo Lahti[3], Miroslava Nedyalkova[4], Ilze Elbere[5], Gennady

ML4 | MICROBIOME


OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.
STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH ALGORITHMS!
SIX MONTHS LATER:
WOW, THIS PROBLEM IS REALLY HARD.
YOU DON'T SAY.

Open workflows

transparency & reproducibility

collaborative research & training

quality & efficiency

ml4microbiome.eu