About the course

- Learning goals
- Course format
- Course materials
- Communication channels
- Introductions

http://web.stanford.edu/class/bios221/Pune/index.html

Learning goals

- Statistical thinking in microbiome studies
- The concept of open and reproducible research
- Familiarity with standard tools in amplicon profiling
- Looking at your own research problems in new ways
- Networking & collaboration!

Statistical Methods in Microbiome Research

Day 1: Taxonomic profiling and analysis

Day 2: Statistical thinking and visualization

Day 3: Population studies and latent variable models

Day 4: Longitudinal analysis

Day 5: Future perspectives

http://web.stanford.edu/class/bios221/Pune/index.html

Course format

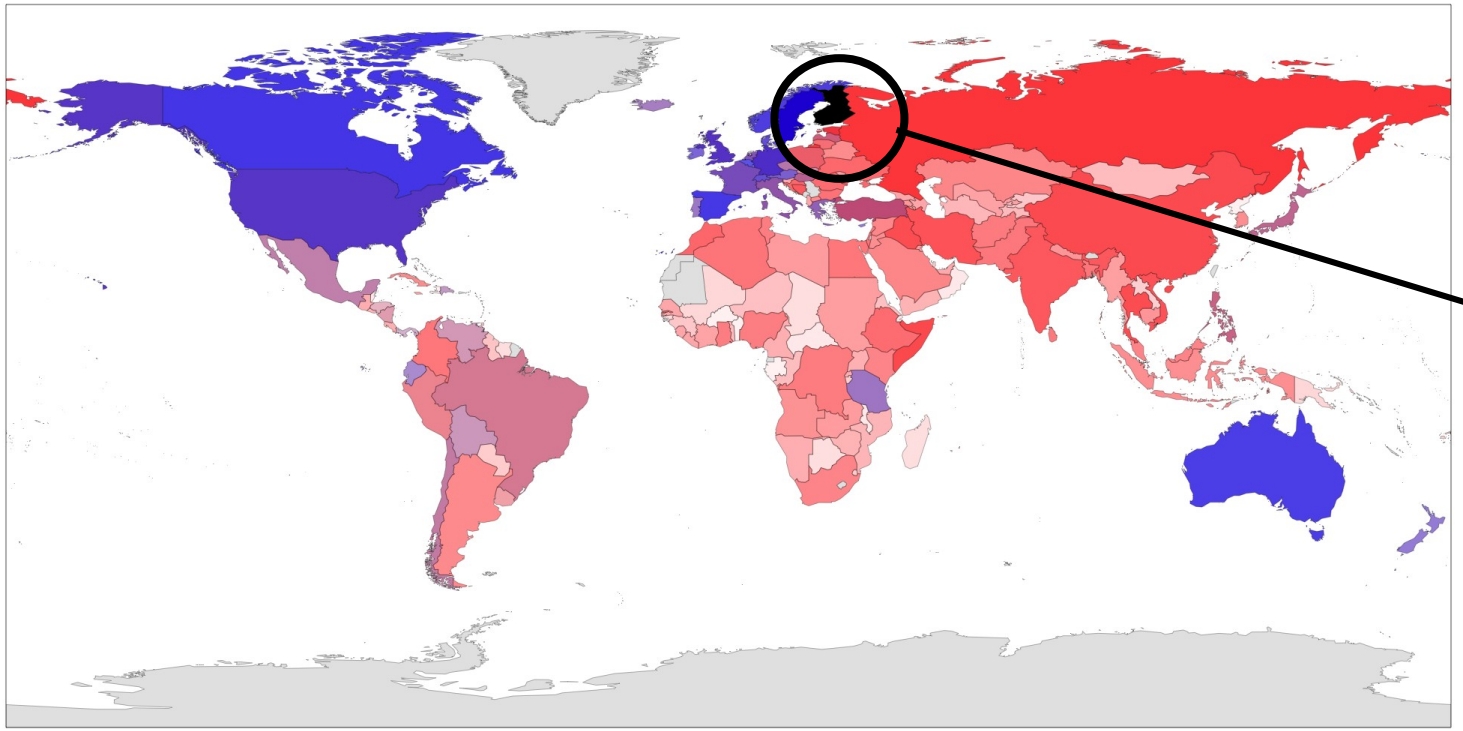
- Active hands-on learning
- Materials available during and after the course
- Mixture of lectures, tutorials, and practice
- Interactive – ask questions & discuss!
- Team learning
- Feedback

Communication channels

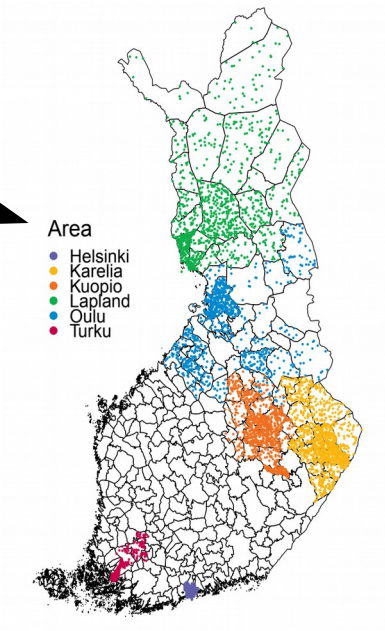
- Website: web.stanford.edu/class/bios221/Pune/index.html
- (Silent) WhatsApp group
- Slack: https://sdacrew.slack.com

  (announcements, links, conversations)
- Twitter: #pune_microbiome_2019
- Color stickers

Introductions



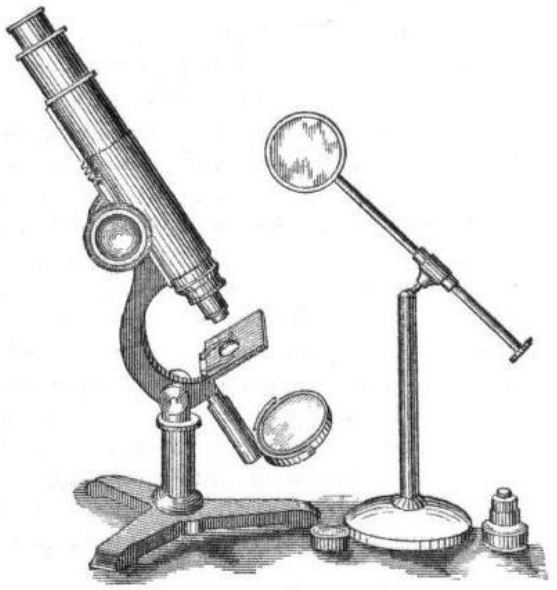
Finland




TURUN YLIOPISTO
VAPAAN KANSAN LAHJA
VAPAALLE TIETEELLE
UNIVERSITY
OF TURKU

Now let’s start!

Antonie van Leeuwenhoek started to investigate **human microbes** ("*animalcules*") around 1670 in the Golden Age of Dutch Science and technology.



Pioneering work in microscopy (275 / 500x) and contributions toward the establishment of microbiology as a scientific discipline.

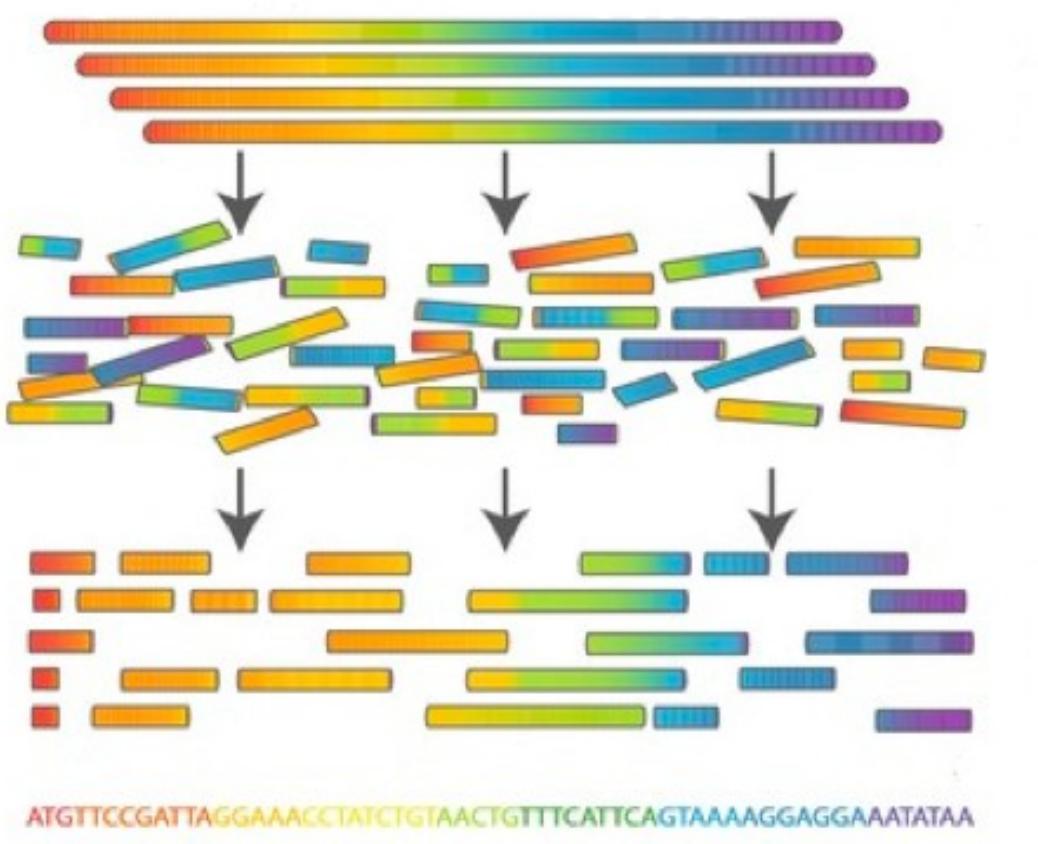
How do we measure microbiome?

Culture-based

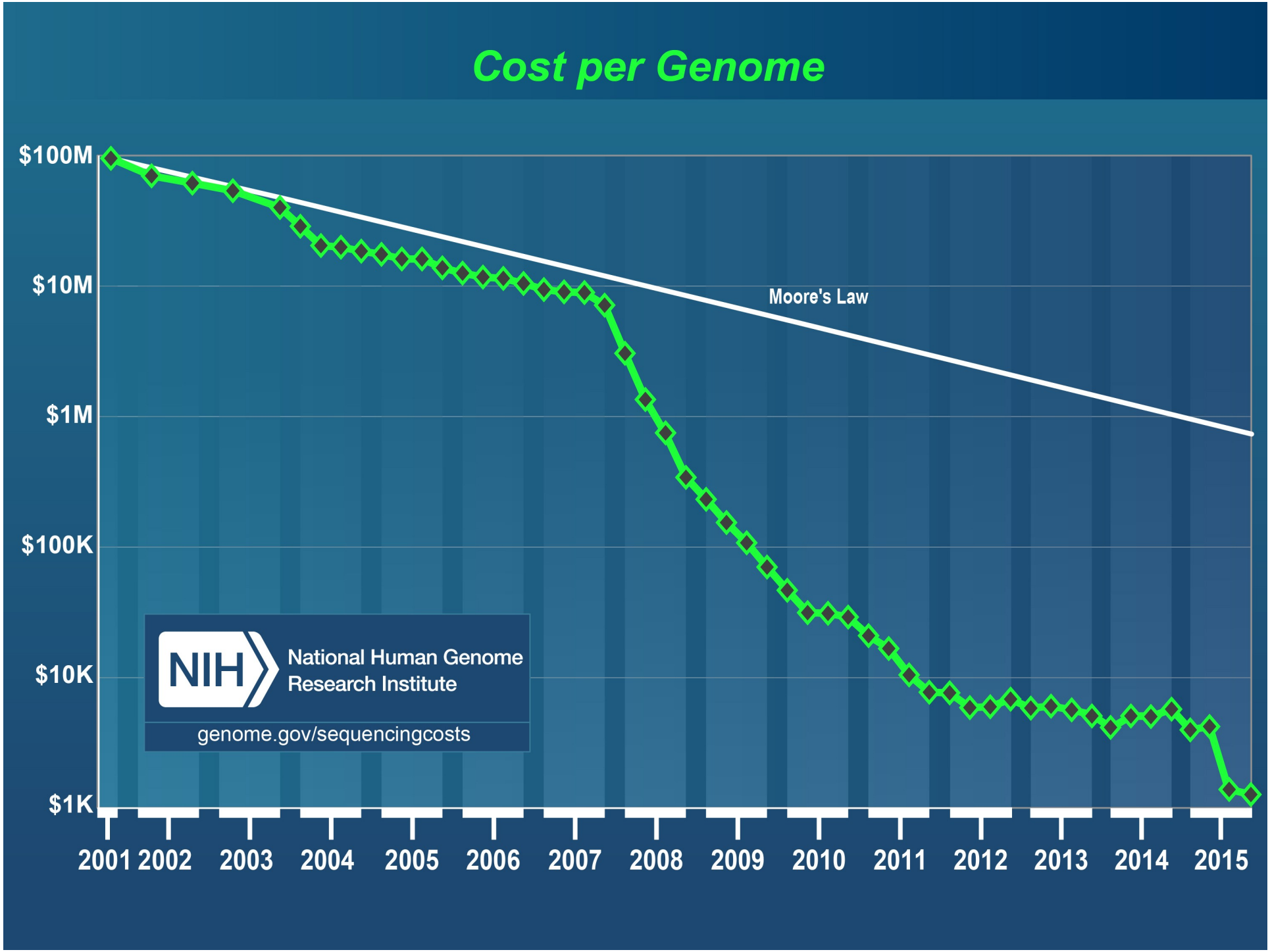
(for *culturable* bugs)

https://www.sott.net/article/309408-A-childs-bacteria-filled-handprint-reveals-the-wonder-of-the-human-microbiome

Sequencing-based
(for *all* bugs; most!)

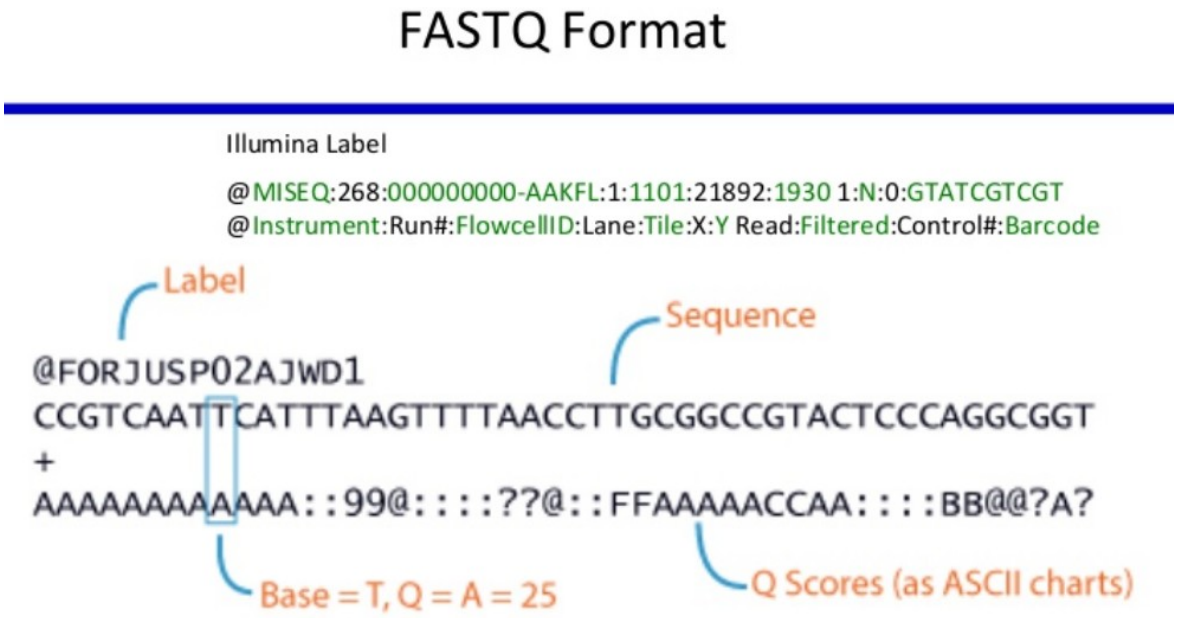


Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. Trends in Ecology and Evolution 29(1): 51-63


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

Standard workflow in microbiome data science

- Raw reads: data retrieval and quality control
- Preprocessing
- Exploration
- Analysis & modeling
- Reproducible reporting



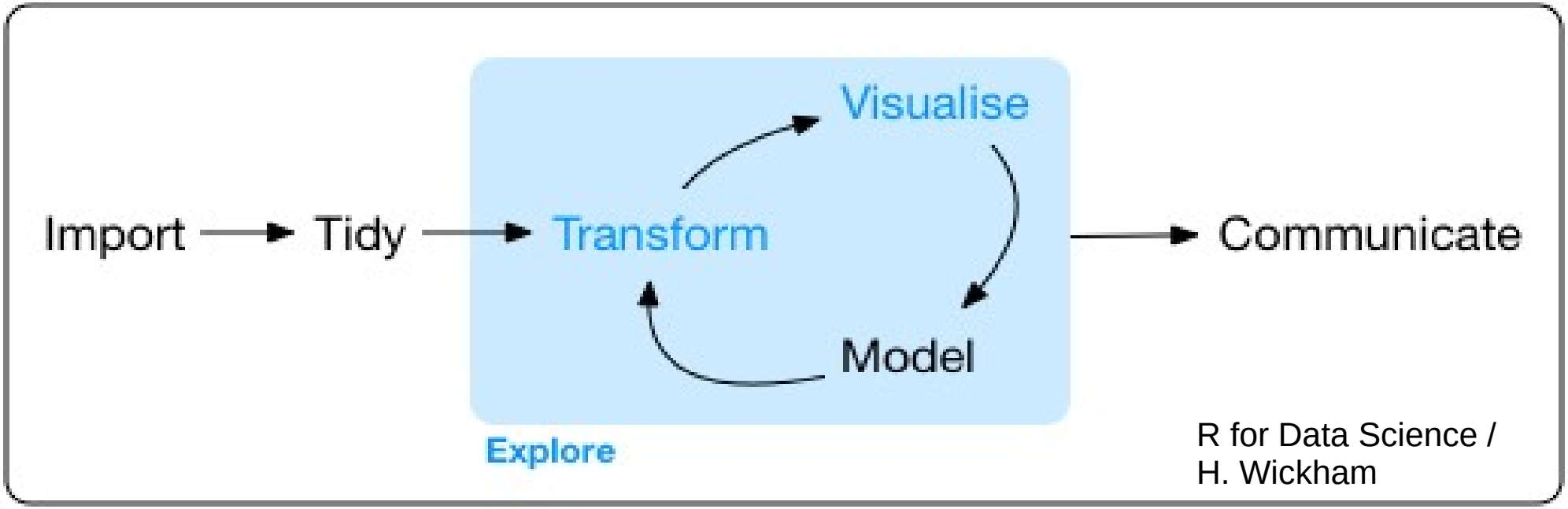
REVISED

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[1]

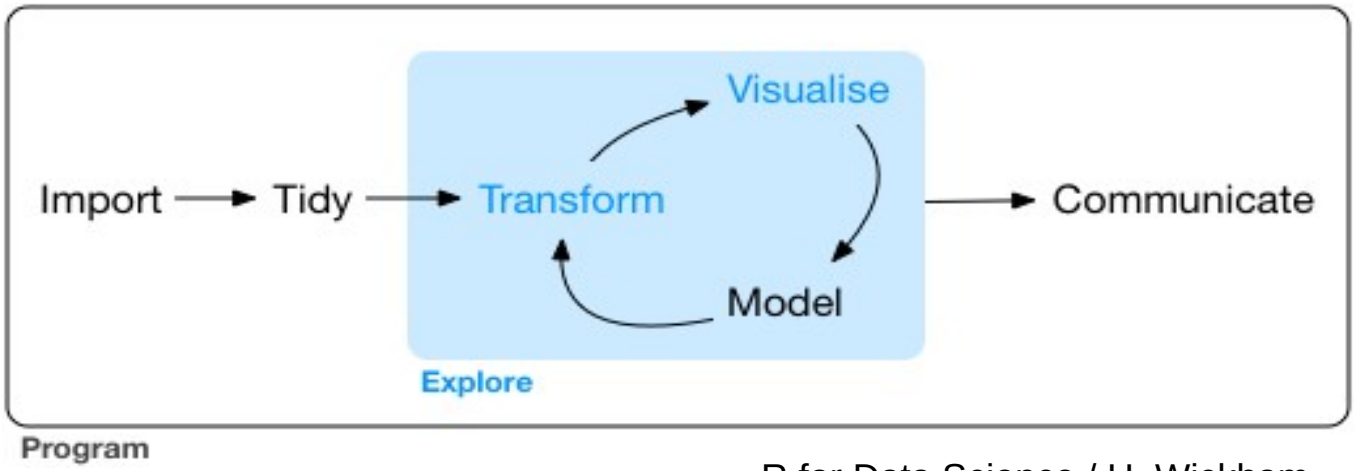
Author details

This article is included in the Bioconductor gateway.



Standard workflow in microbiome data science

- Raw reads: data retrieval and quality control
- Preprocessing
- Exploration
- Analysis
- Modeling
- Reproducible reporting



R for Data Science / H. Wickham

FASTQ Format

Illumina Label

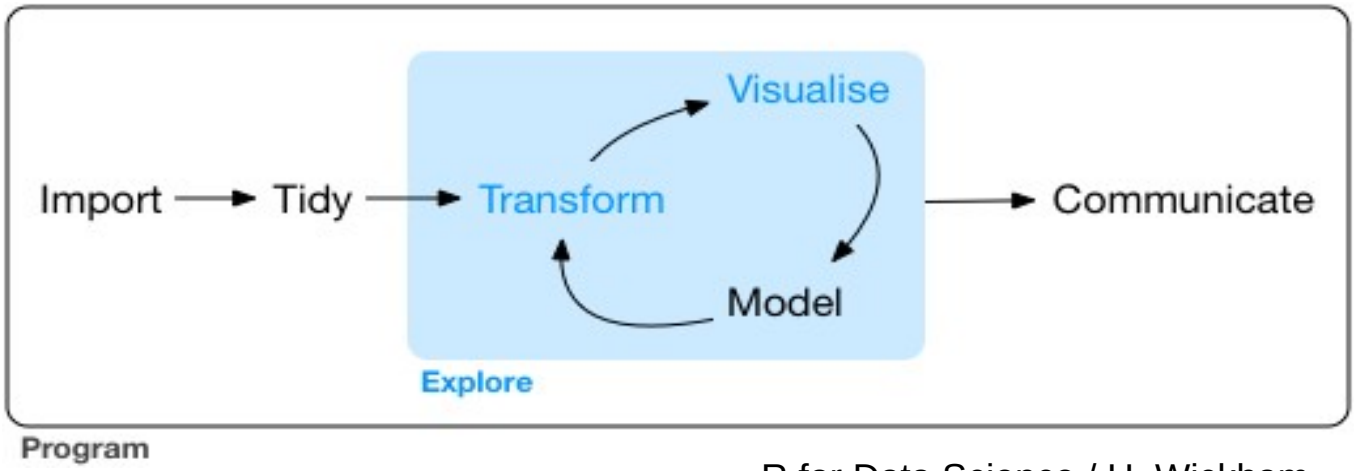
@MISEQ:268:000000000-AAKFL:1:1101:21892:1930 1:N:0:GTATCGTCGT
@Instrument:Run#:FlowcellID:Lane:Tile:X:Y Read:Filtered:Control#:Barcode

Once demultiplexed fastq files without non-biological nucleotides are in hand, the dada2 pipeline proceeds as follows:

1. Filter and trim: `filterAndTrim()`
2. Dereplicate: `derepFastq()`
3. Learn error rates: `learnErrors()`
4. Infer sample composition: `dada()`
5. Merge paired reads: `mergePairs()`
6. Make sequence table: `makeSequenceTable()`
7. Remove chimeras: `removeBimeraDenovo()`

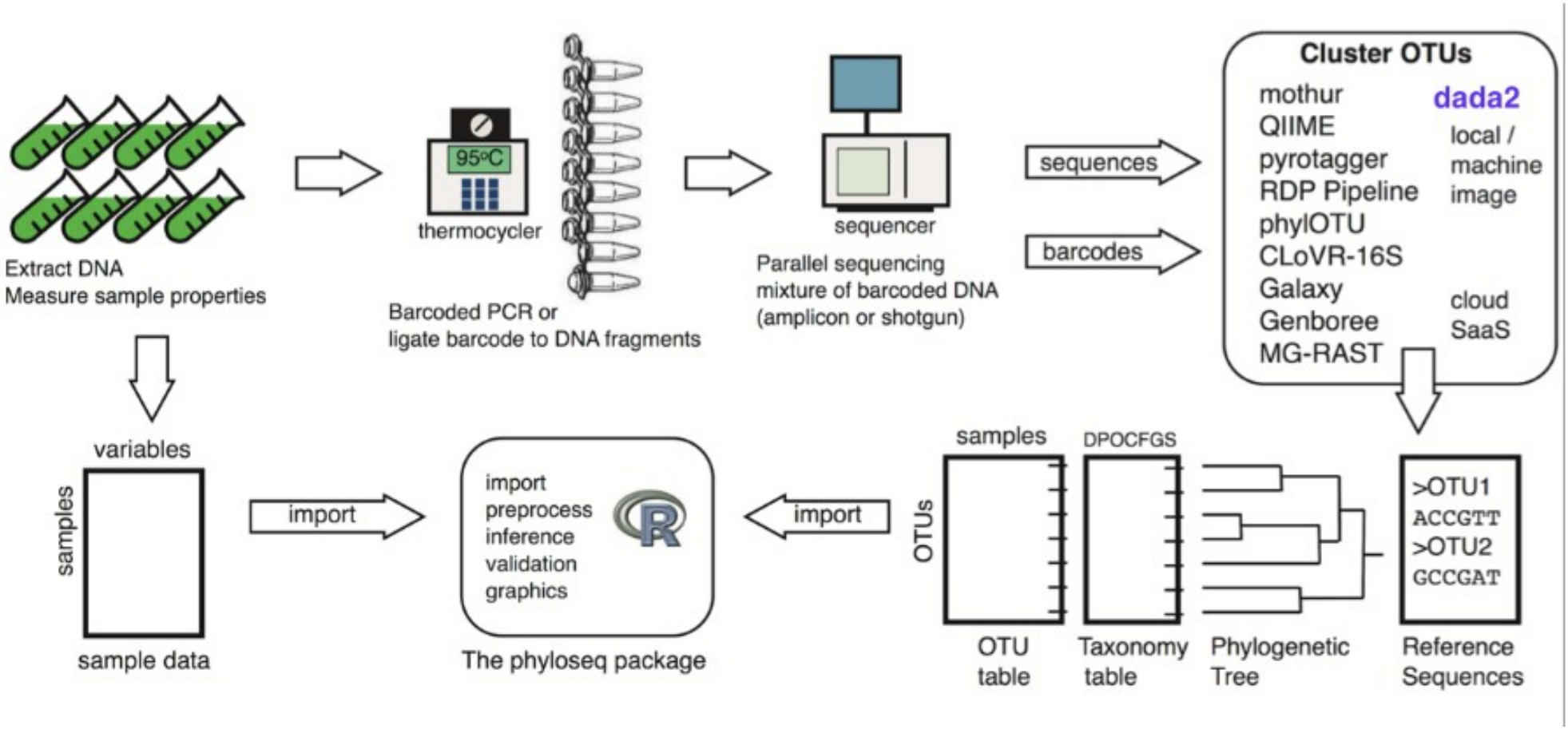
Standard workflow in microbiome data science

- Raw reads: data retrieval and quality control
- Preprocessing
- Exploration
- Analysis & modeling
- Reproducible reporting



R for Data Science / H. Wickham

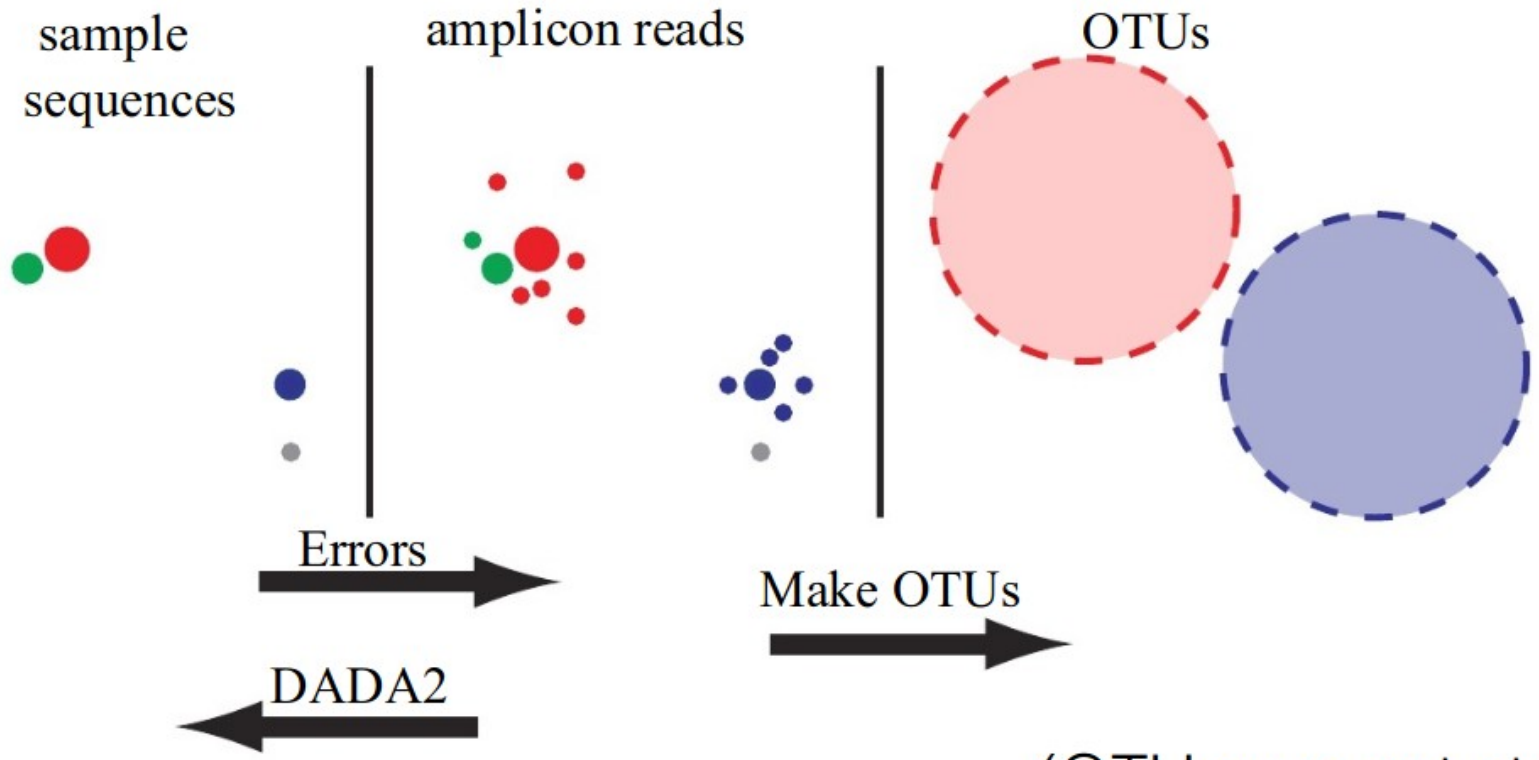
Phyloseq



(McMurdie and Holmes, 2013, Plos ONE).

Source: Susan Holmes | http://web.stanford.edu/class/bios221/Short-Phyloseq-Resources.html

Goal: Infer original sequences from noisy reads



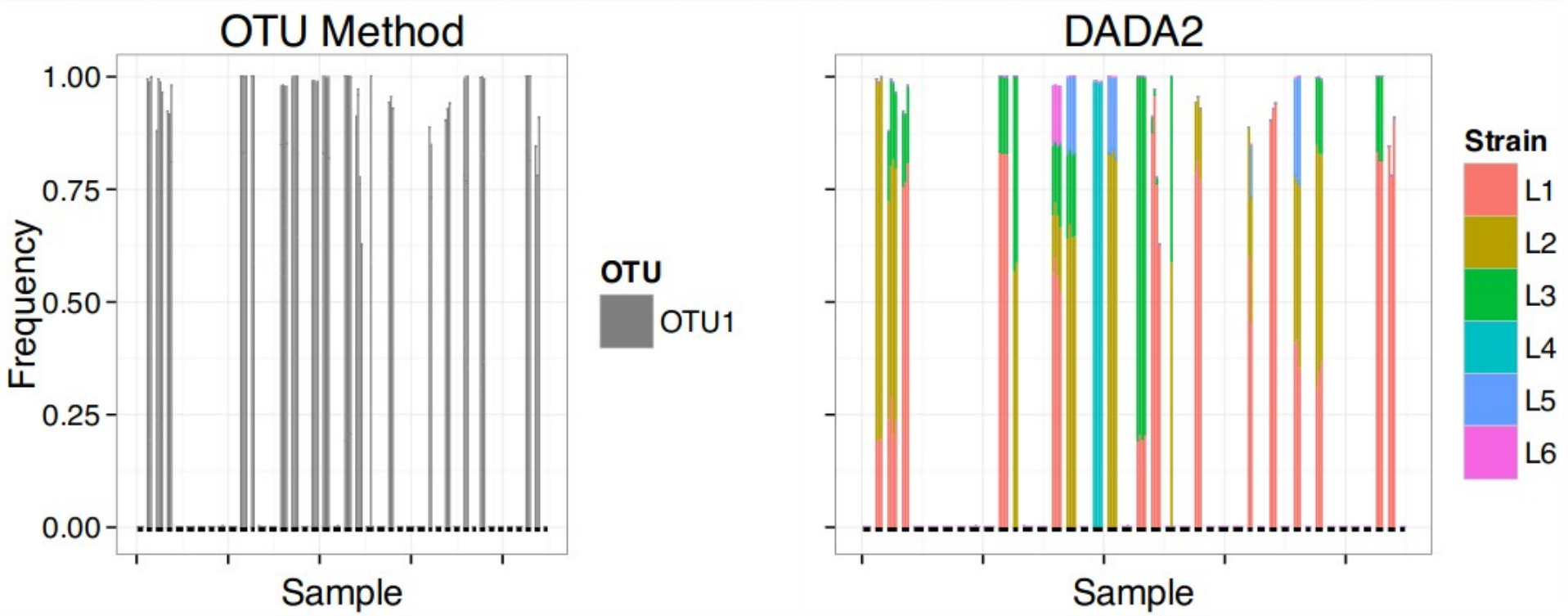
(OTUs are not strains)

OTUs: Lump similar sequences together
DADA2: Statistically infer the sample sequences (strains)

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Real example, exact sequence resolution

Lactobacillus crispatus sampled from
vaginal microbiome 42 pregnant women



Data: MacIntyre et al. Scientific Reports, 2015.

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Perspective

Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

OPEN

Benjamin J Callahan[1], Paul J McMurdie[2] and Susan P Holmes[3]

[1]Department of Population Health and Pathobiology, NC State University, Raleigh NC, USA
[2]Whole Biome Inc, San Francisco CA, USA
[3]Department of Statistics, Stanford University, Stanford CA, USA

The ISME Journal 21 July 2017; doi: 10.1038/ismej.2017.119

The ISME Journal
Multidisciplinary Journal of Microbial Ecology

Other relevant articles:

UNOISE2 — *bioRxiv* **Oct 2016** 081257

Deblur — *mSystems* **Mar 2017** 2 (2) e00191-16 Unknown

*MED — *The ISME Journal* **2015** 9, 968–979 High FP!

*DADA1 — *BMC bioinformatics* **2012** 13(1), 283 Slow

PyroNoise — BMC Bioinformatics Quince et al. **2011** 454 only

nature biotechnology

Resource | Published: 27 August 2018

A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil & Philip Hugenholtz

Nature Biotechnology **36**, 996–1004 (2018) | Download Citation

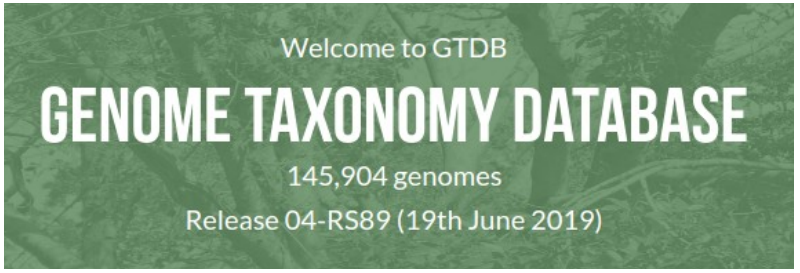
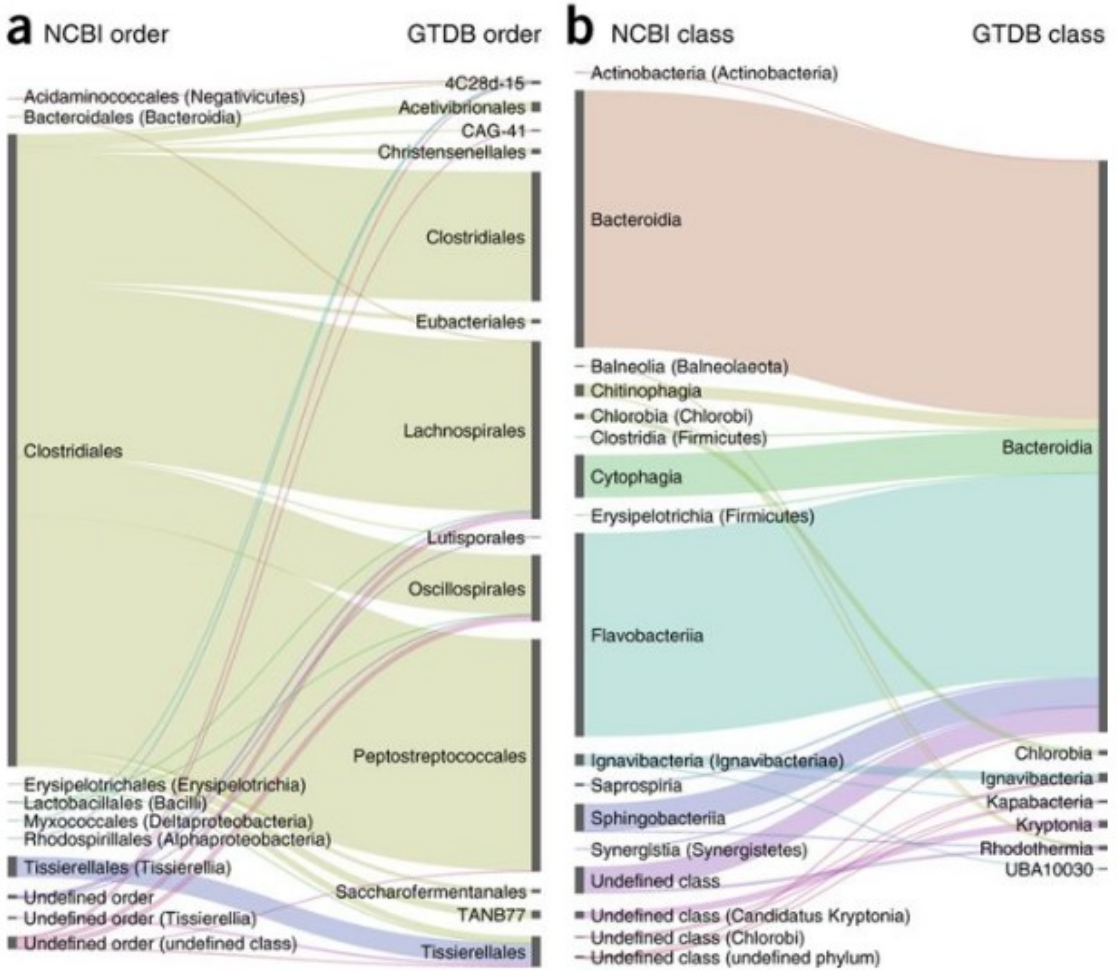


Figure 5: Comparisons of NCBI and GTDB classifications of genomes designated as Clostridia or Bacteroidetes in the GTDB taxonomy.



Cold Spring Harbor Laboratory

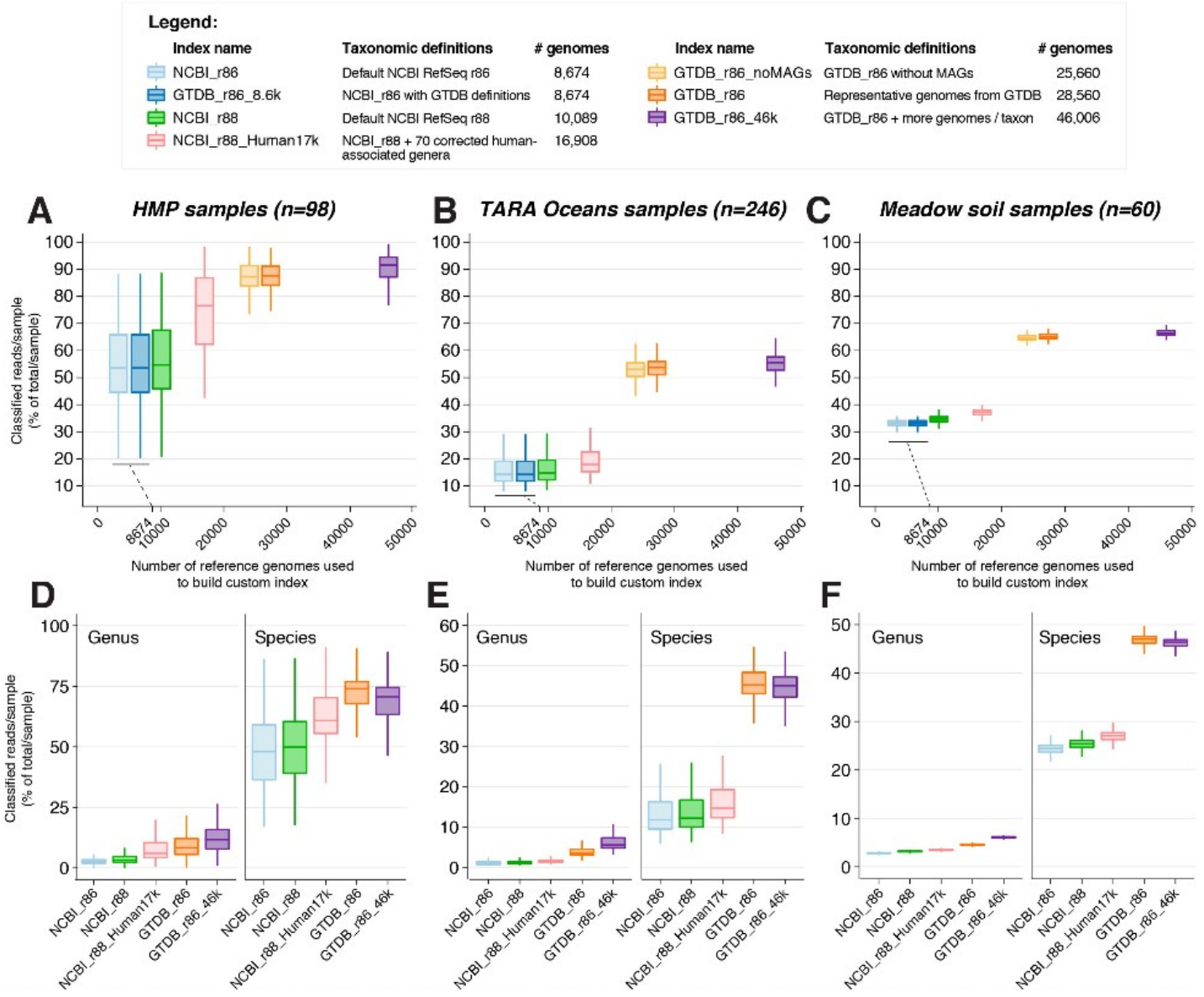
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Correcting index databases improves metagenomic studies

Guillaume Méric, Ryan R. Wick, Stephen C. Watts, Kathryn E. Holt, Michael Inouye

doi: https://doi.org/10.1101/712166

GTDB can substantially improve metagenomic read classification and analyses



Data organization

Side information on samples

Metadata
(age, bmi, sex..)

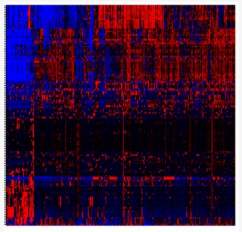
Side information on taxonomic units

Ref Seq

Phylo tree

Tax table

OTU Abundances

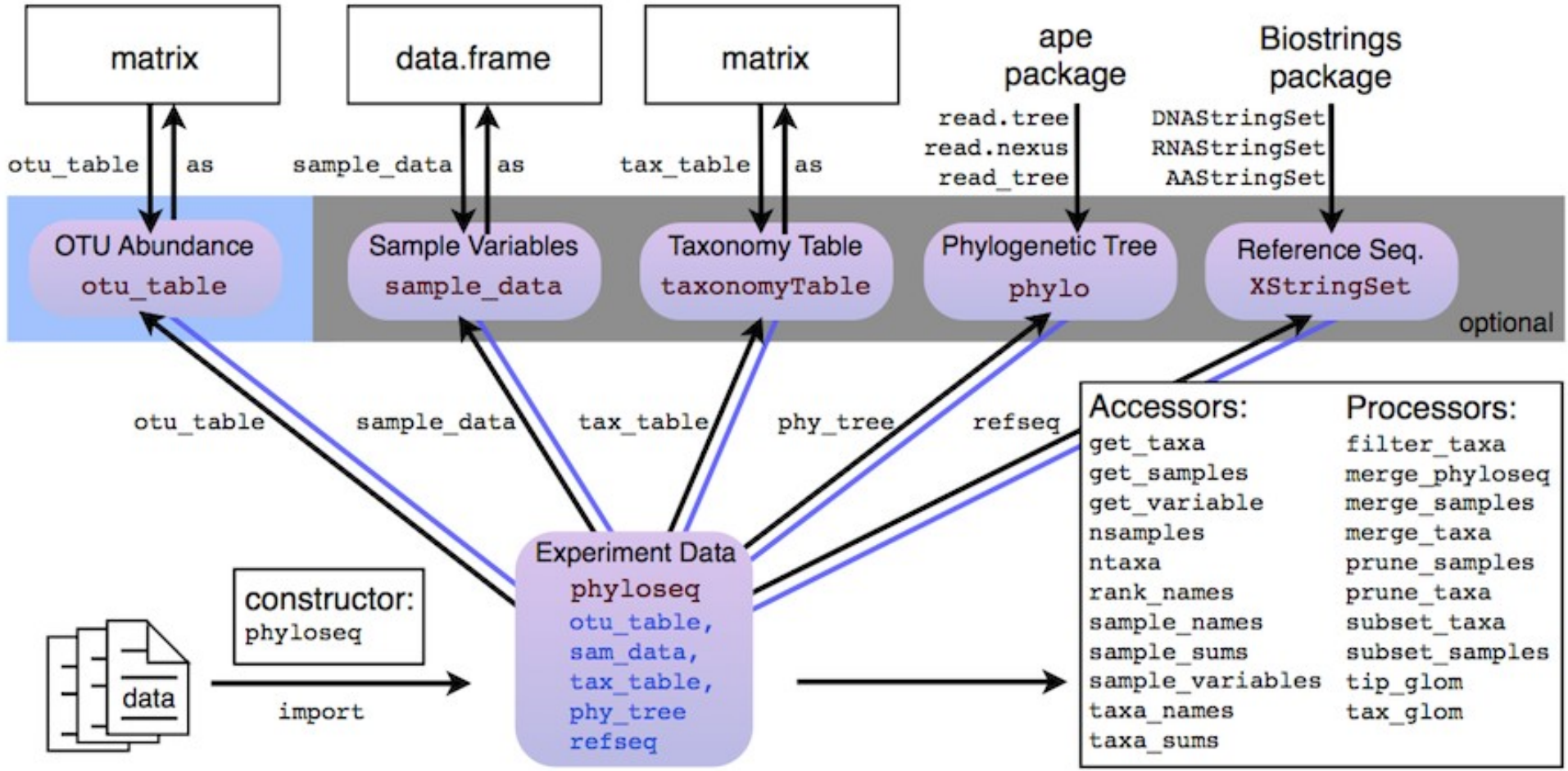


OTUs

Samples

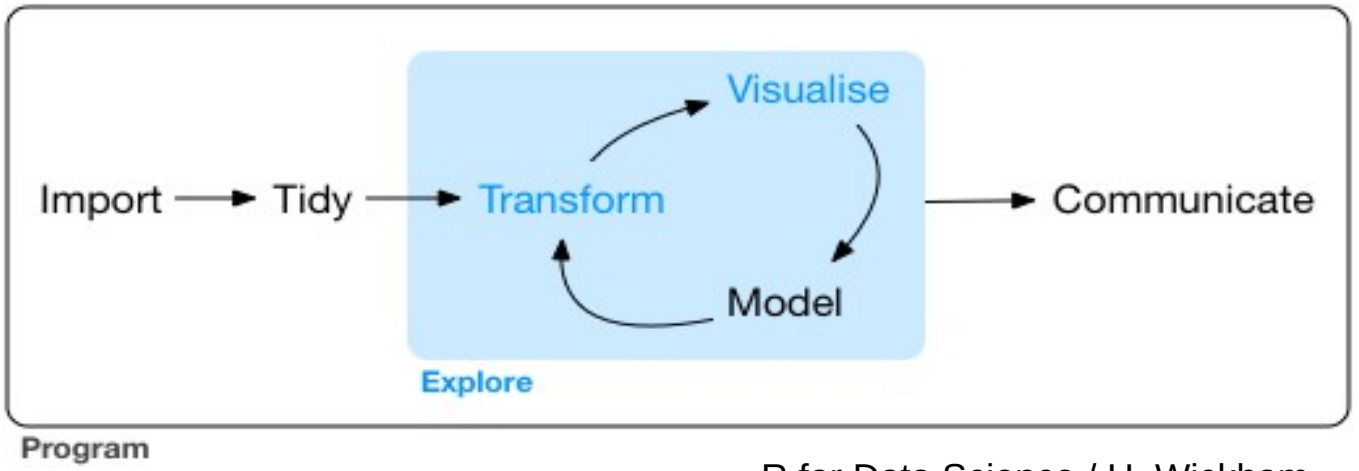
Data standard: *phyloseq*

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



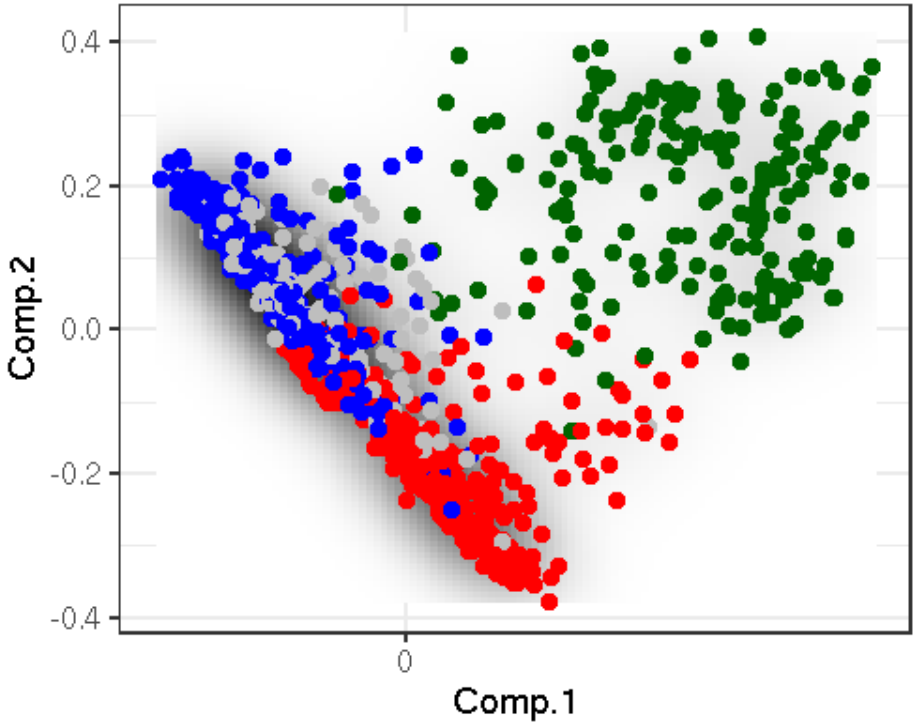
Standard workflow in microbiome data science

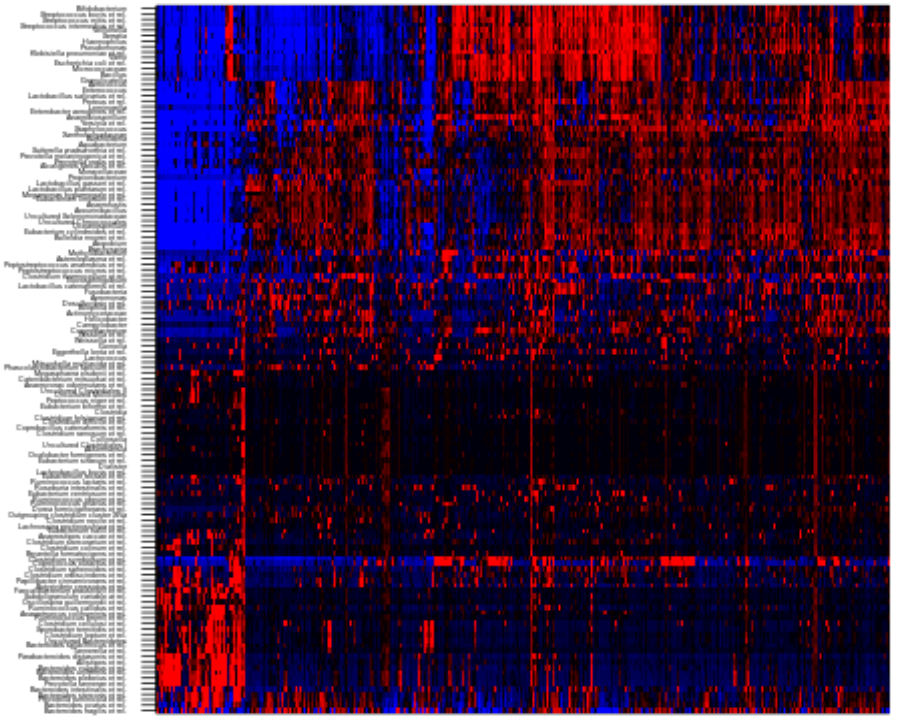
- Raw reads: data retrieval and quality control
- Preprocessing
- Exploration
- Analysis & modeling
- Reproducible reporting



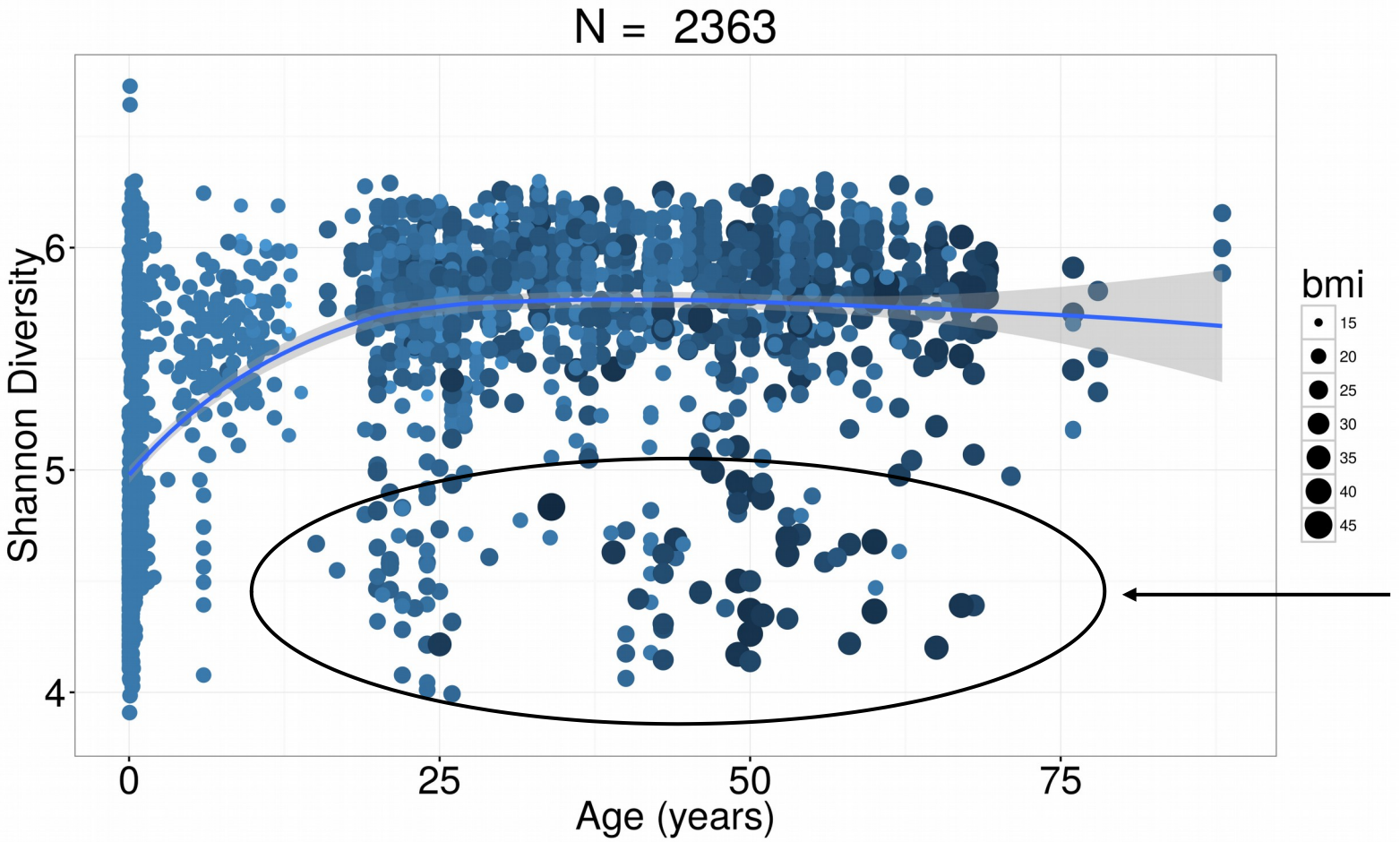
R for Data Science / H. Wickham

Visual exploration, unsupervised analysis





Aging, microbiome diversity & tipping elements: healthy & normal obese subjects



Confounding variables: stool consistency showed the largest effect size on microbiota variation

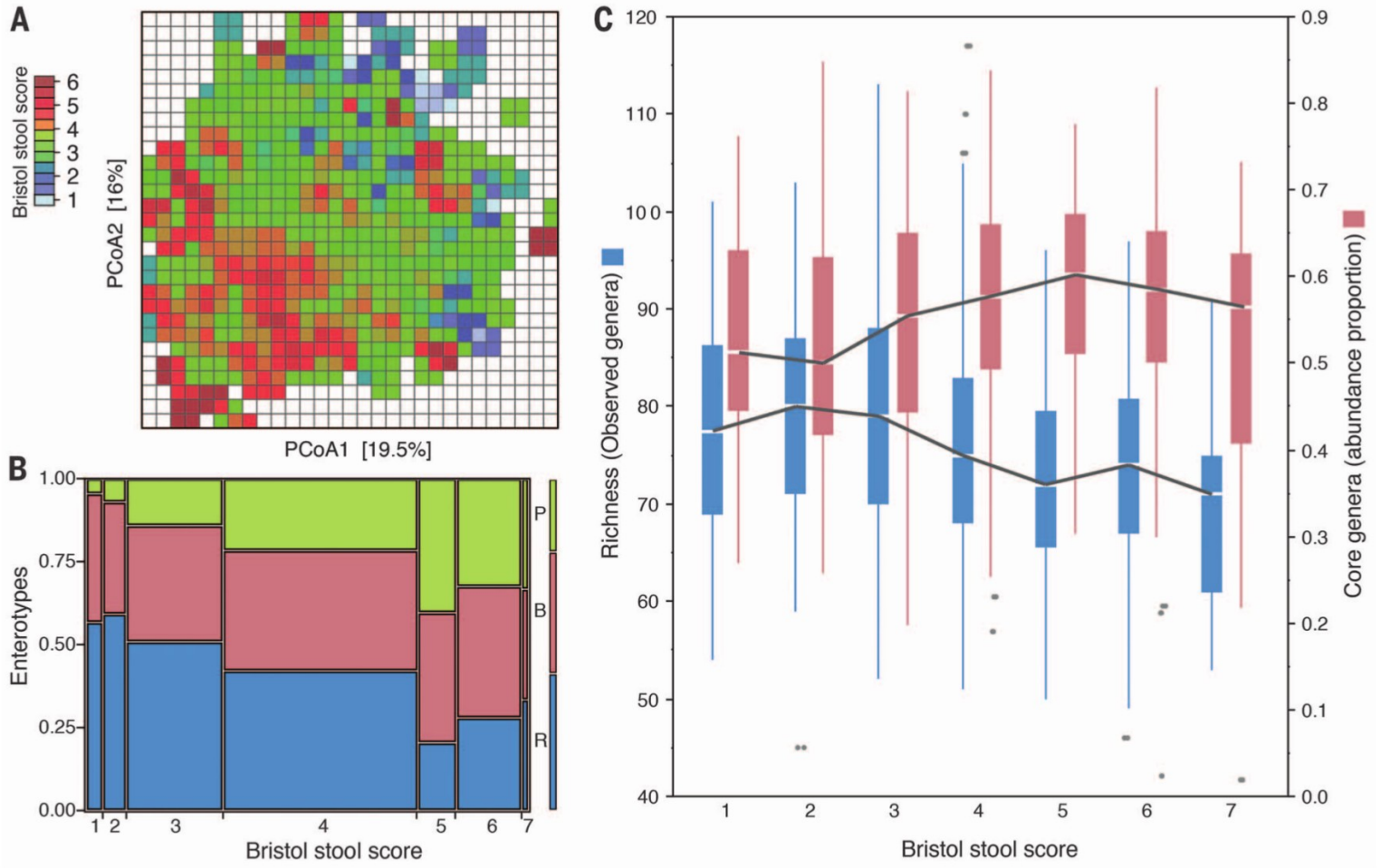


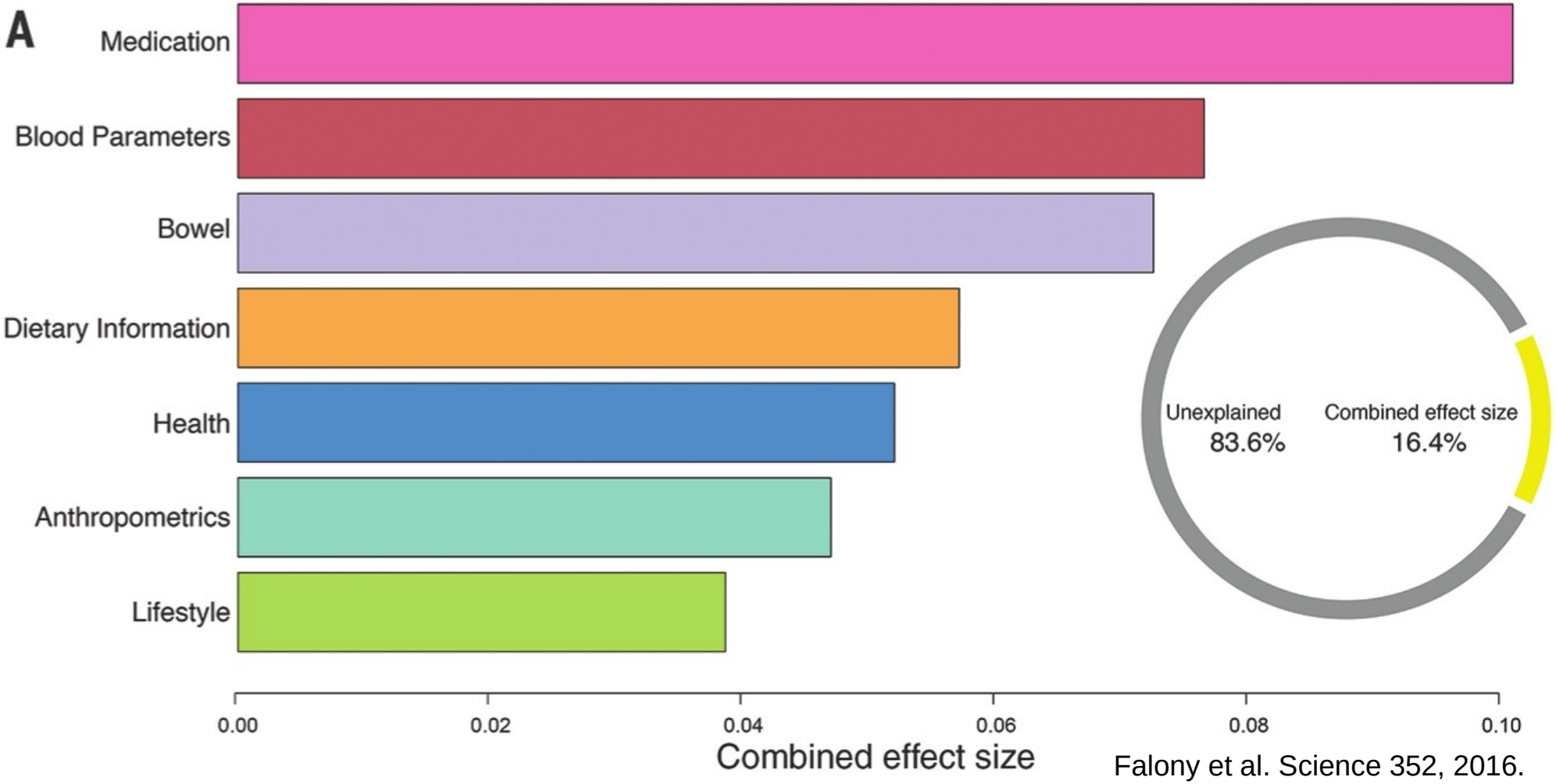
Fig. 4. BSS score association to microbiota variation. (**A**) BSS score variation across the FGFP cohort, as represented on the genus-level PCoA ordination (Bray-Curtis dissimilarity). Each cell is colored according to median BSS score of individual samples allocated to the cell coordinates. (**B**) Enterotype distribution over BSS scores [JSD enterotyping (*18*)] showing an increase in *Prevotella* individuals with looser stool consistency. (**C**) Median differences in abundance of the core microbiota (FGFP genus-level core at 99%) and in observed genus richness across BSS score.

Falony et al. Science 352, 2016.

Total explained variation: 16.4%

(Flemish Gut Flora Project)

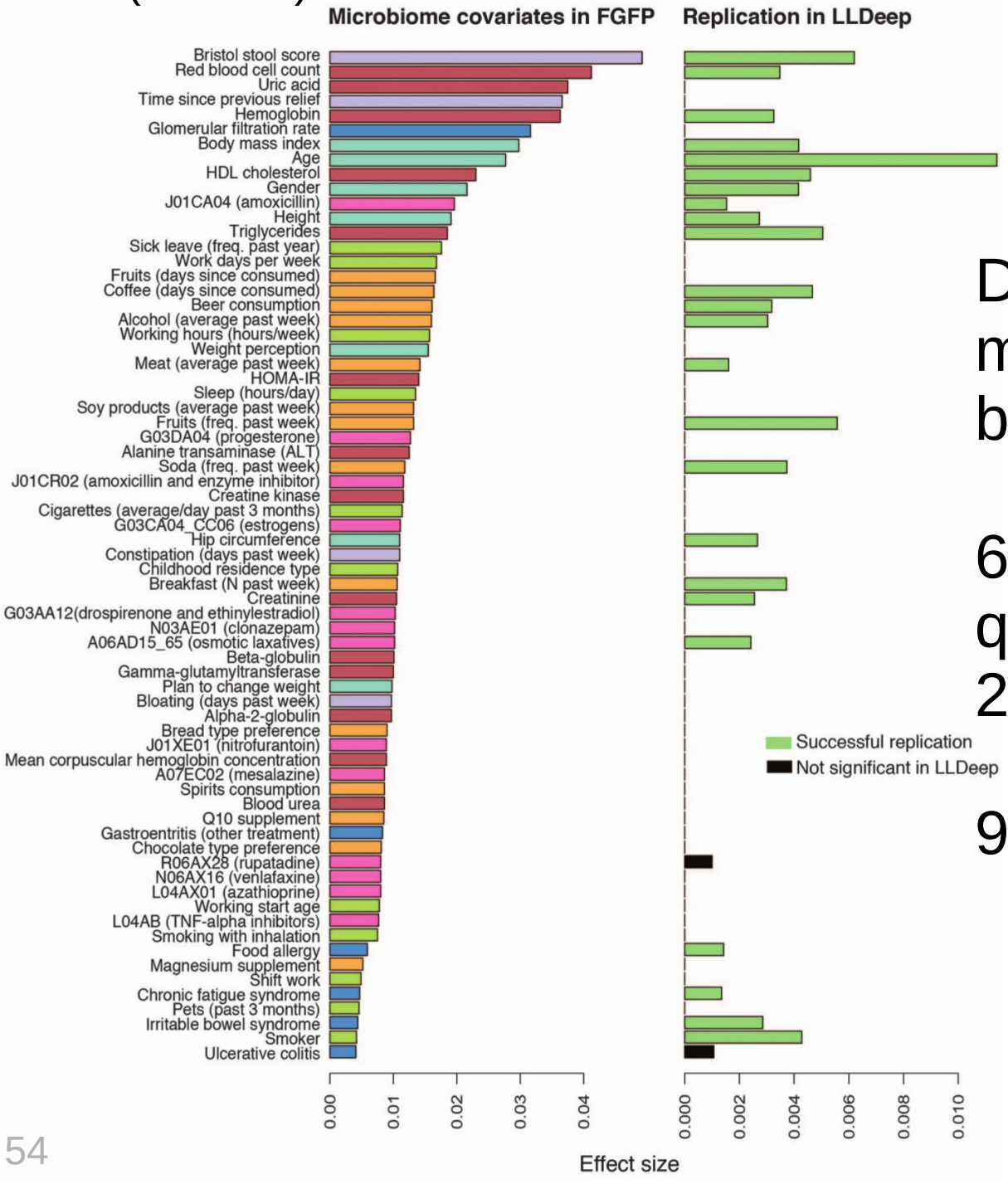
Proposed disease marker genera associated to host covariates and medication - inclusion in study design is essential !



Falony et al. Science 352, 2016.

Flemish Gut Flora (N=1106)

Dutch LifeLines-DEEP (N=1135)



Reproducibility (experimental & computational)?

Diet, health, lifestyle, medication, host variables, blood, bowel habits..

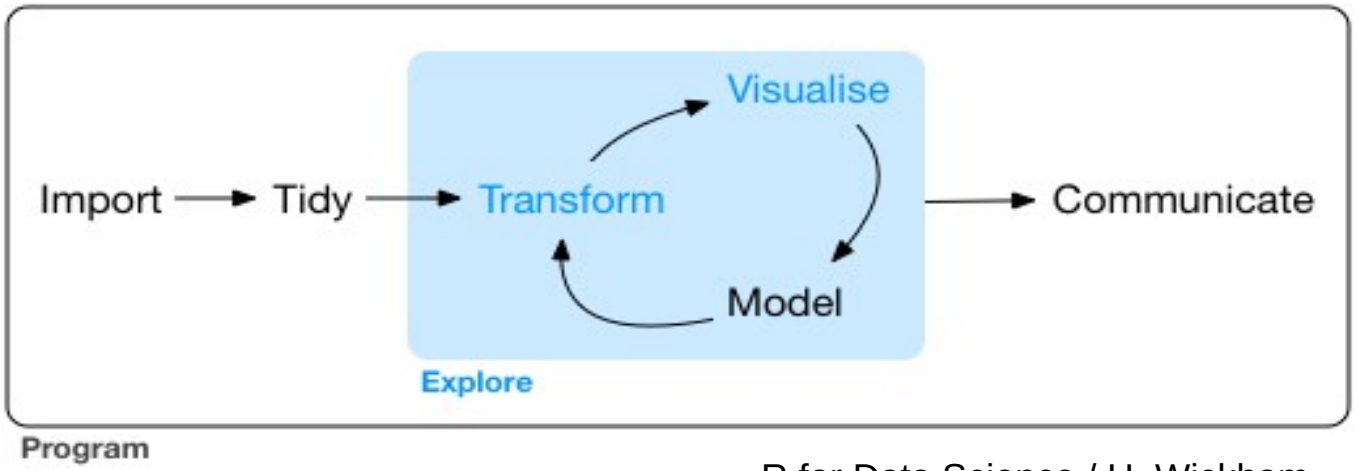
69 covariates (clinical & questionnaire) in FGFP; 26 shared with DEEP

92% replication rate !

Falony et al. Science 352, 2016.

Standard workflow in microbiome data science

- Raw reads: data retrieval and quality control
- Preprocessing
- Exploration
- Analysis & modeling
- Reproducible reporting



R for Data Science / H. Wickham

Associations?

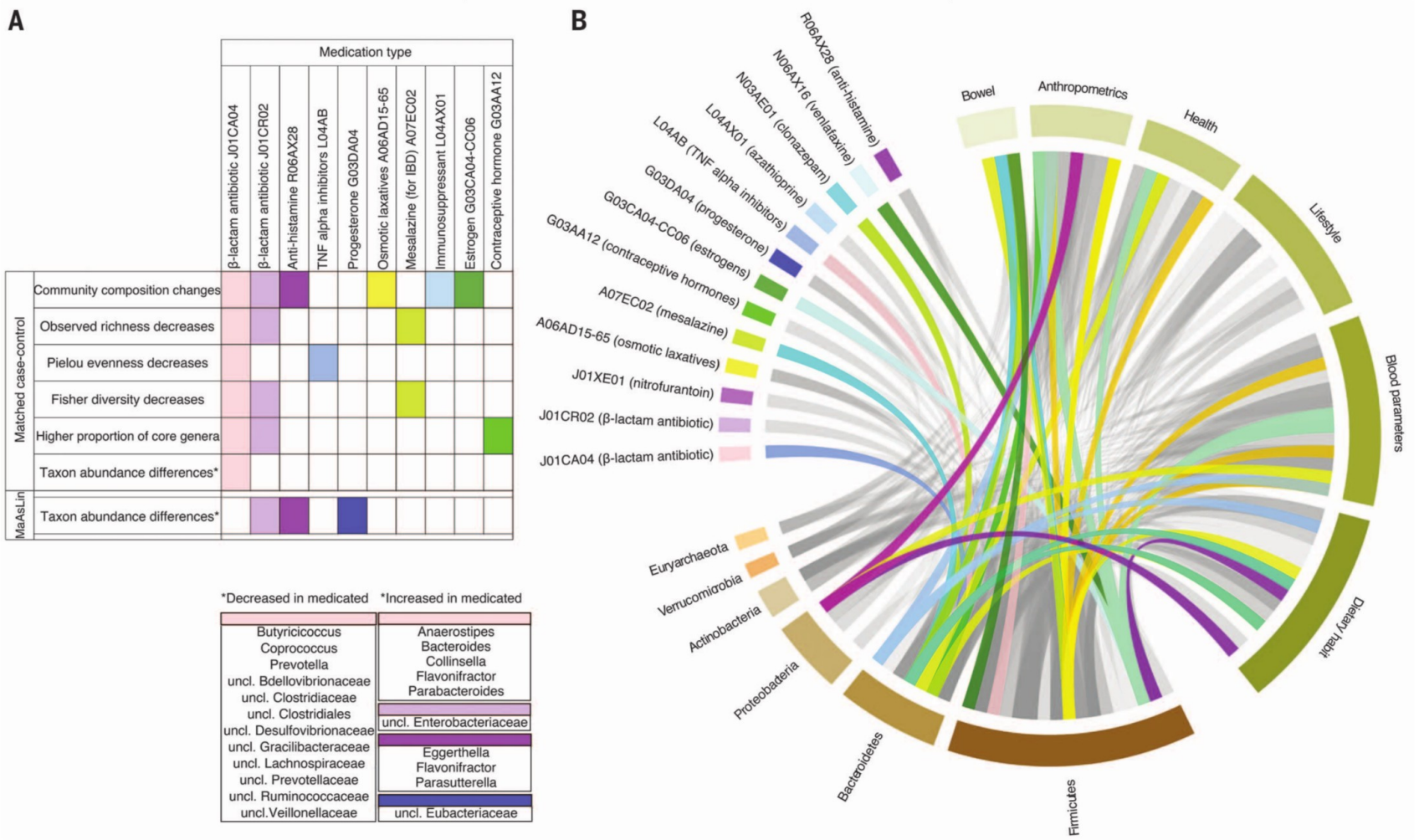
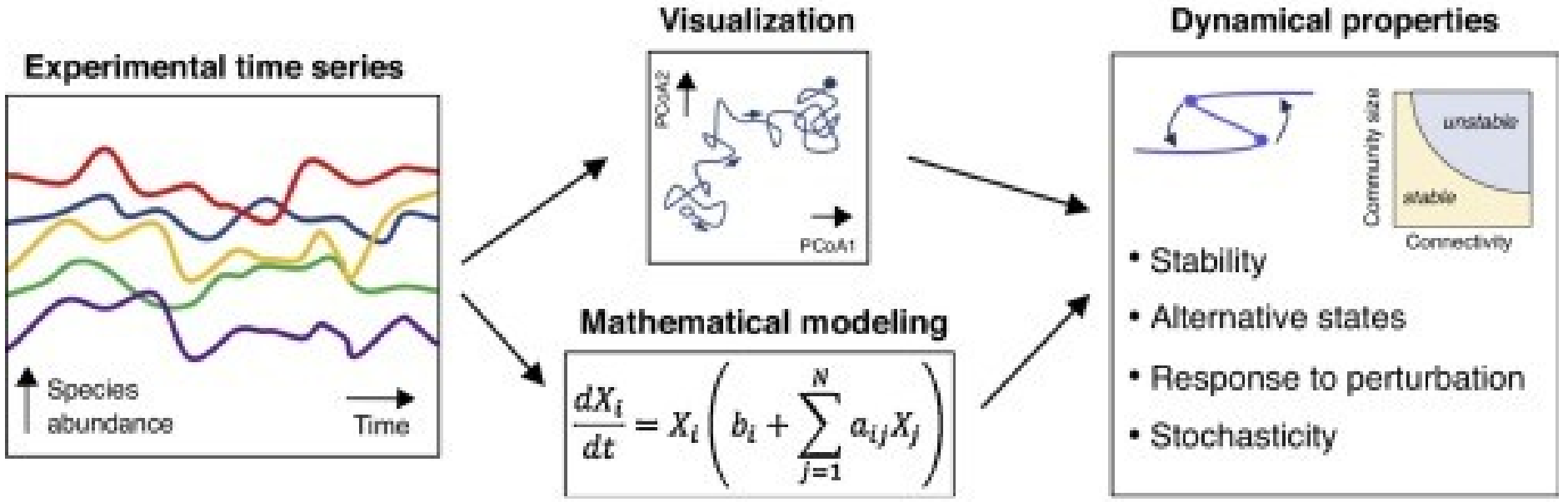


Fig. 5. Drug interactions in the FGFP. (**A**) Overview of the association between different types of medication and microbiome composition. Colored boxes (color coding according to medication) represent a significant result in the matched case-control (FDR<5%) or boosted additive general linear modeling (FDR<10%, table S11) analyses. The effect (decrease/increase) of medication on genera abundances is specified. (**B**) Circos plot showing correlations between covariates and genus abundances (FDR<10%) interacting with drugs. Genera are grouped at phylum level; ribbons represent genus-phenotype associations and are colored according to the confounding medication (gray indicates nonconfounded).

Mechanisms

Microbial communities as dynamical systems

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



Community assembly: mixture of ecological processes:
combinations of mechanistic & non-parametric models needed

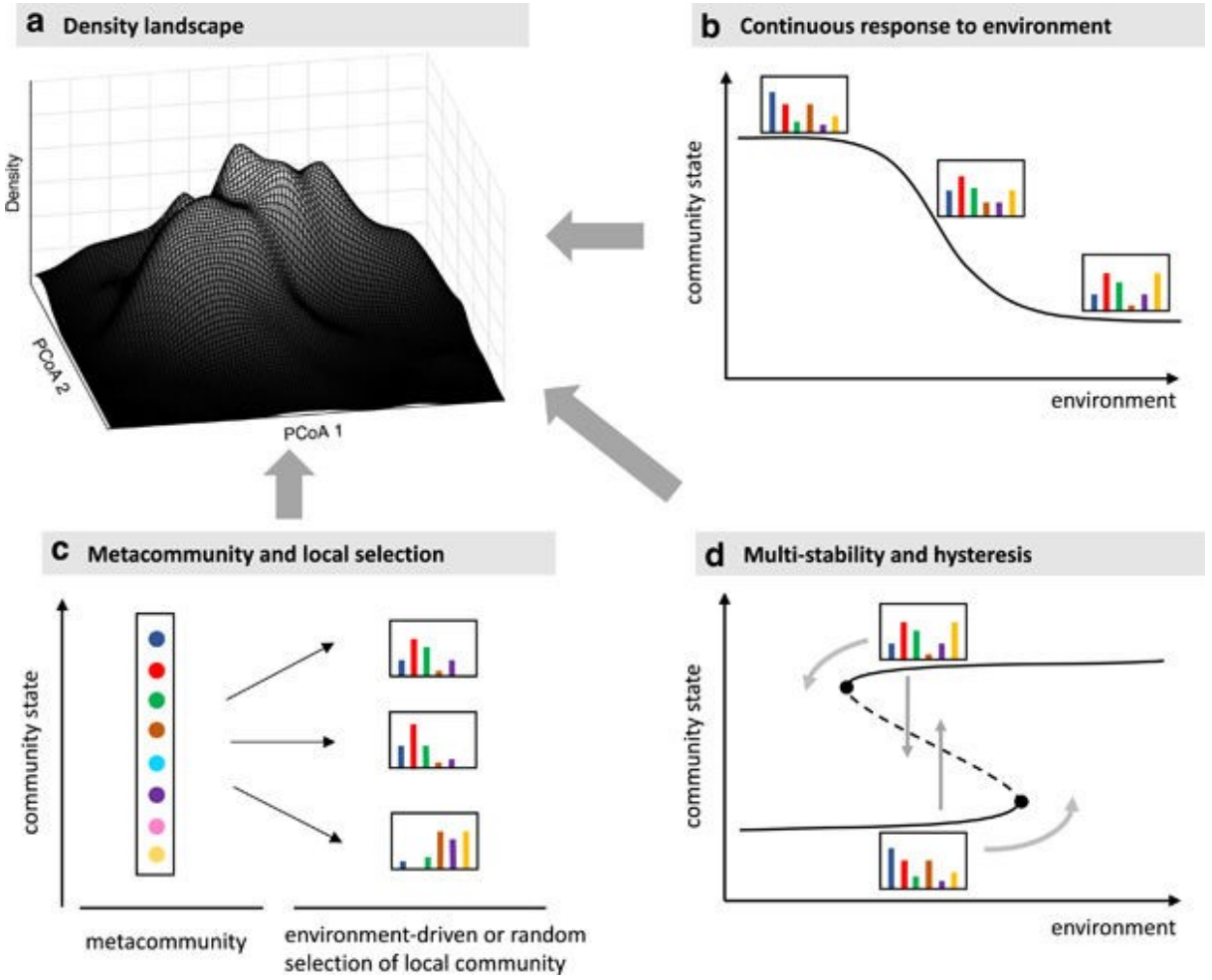
Multi-stability and the origin of microbial community types

Didier Gonze, Leo Lahti, Jeroen Raes & Karoline Faust

The ISME Journal (2017) **11**, 2159–2166 (2017)
doi:10.1038/ismej.2017.60
Download Citation
Microbial communities | Population dynamics
Received: 06 December 2016
Revised: 28 February 2017
Accepted: 10 March 2017
Published online: 05 May 2017

Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process

Keith Harris[1], Todd L Parsons[2], Umer Z Ijaz[3], Leo Lahti[4], Ian Holmes[5], Christopher Quince[6,*]



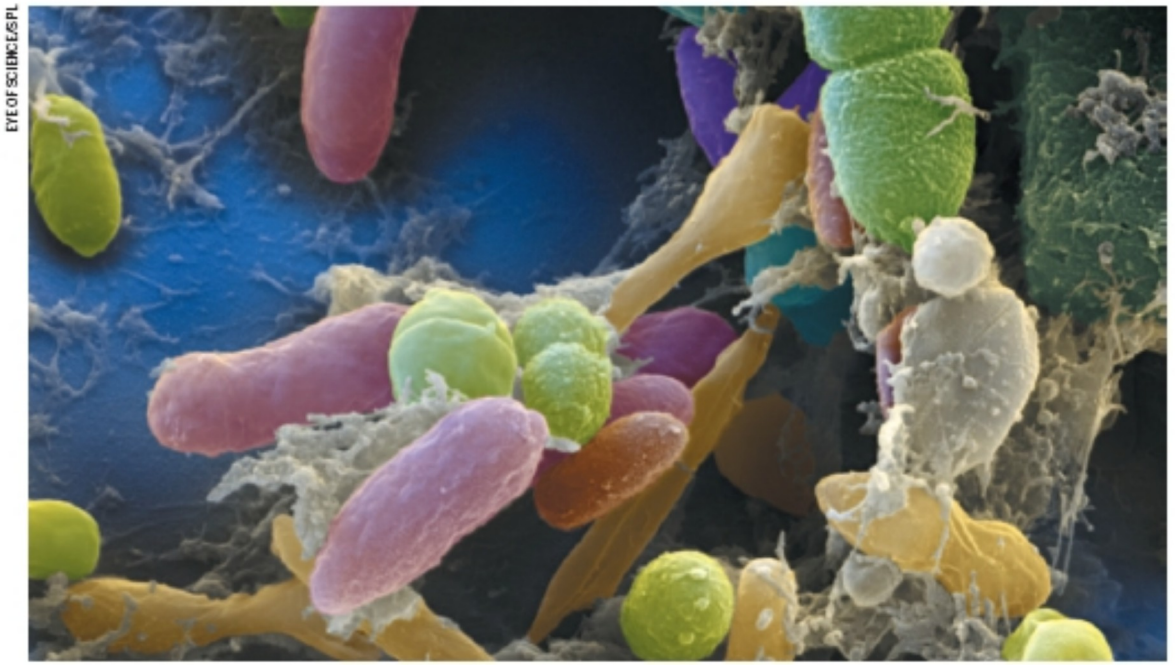
- Alternative states
- Hysteresis
- Periodicity
- Metacommunities
- Niche models
- Competition
- Mutualism

- Environmental drivers

- Host interactions

- Chaos
- Neutral processes & Stochasticity

- etc ...


EYE OF SCIENCE/SPL

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

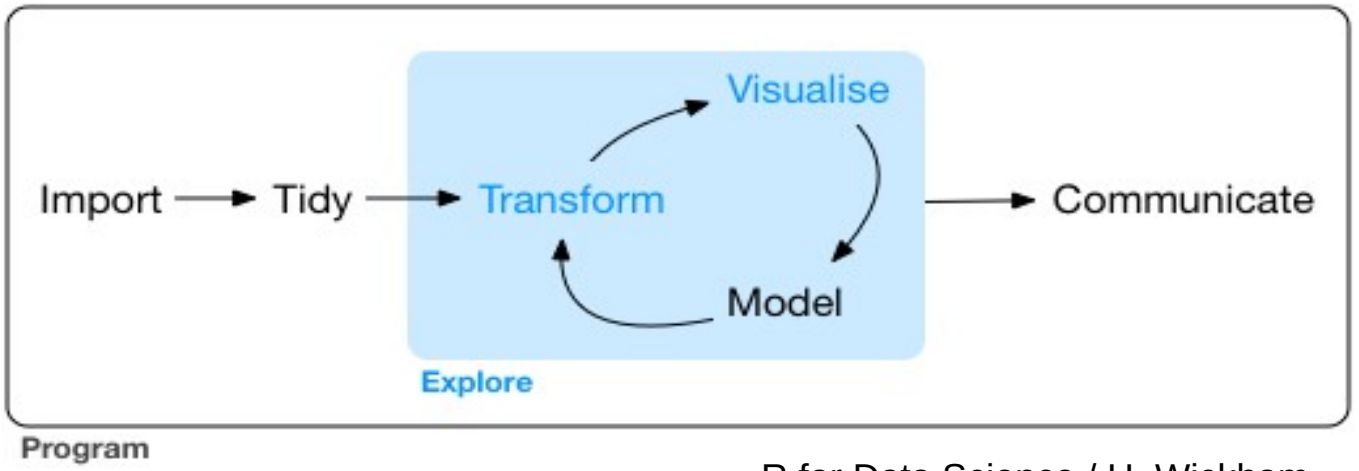
Microbiome science needs a healthy dose of scepticism

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

Standard workflow in microbiome data science

- Raw reads: data retrieval and quality control
- Preprocessing
- Exploration
- Analysis & modeling
- Reproducible research



R for Data Science / H. Wickham

"*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.

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: Aug 30, 2005 • DOI: 10.1371/journal.pmed.0020124

PLOS | MEDICINE

Browse | For Authors | About Us | S

OPEN ACCESS

ESSAY

898,944 VIEWS | 1,119 CITATIONS

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: Aug 30, 2005 • DOI: 10.1371/journal.pmed.0020124

How to Make More Published Research True

John P. A. Ioannidis

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

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,*]

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

Source: Wikimedia Commons / Public domain

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

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

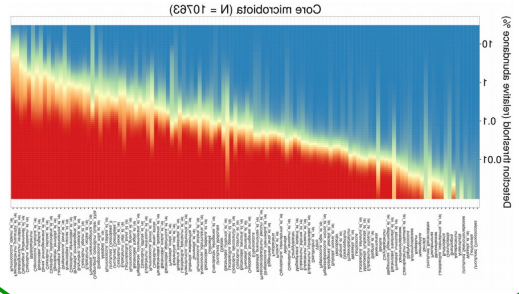
microbiome R package

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

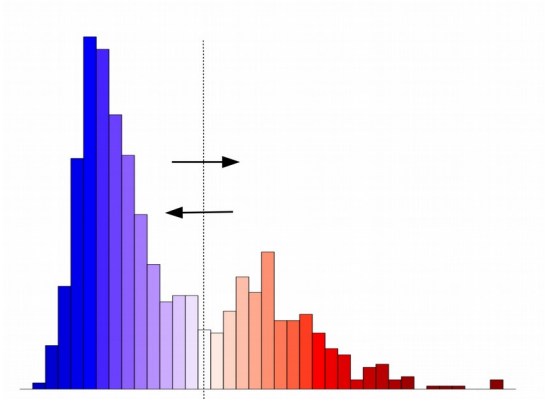
chat on gitter | build passing | codecov 24% | PRs welcome

Core & prevalence

prevalence(x)
core(x)
core_members(x)

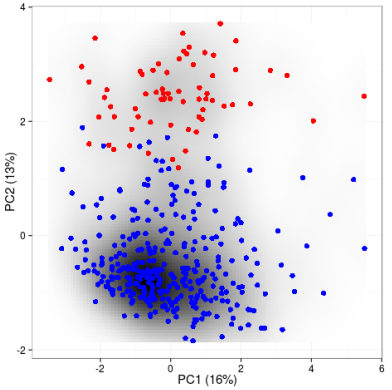


Stability & resilience



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

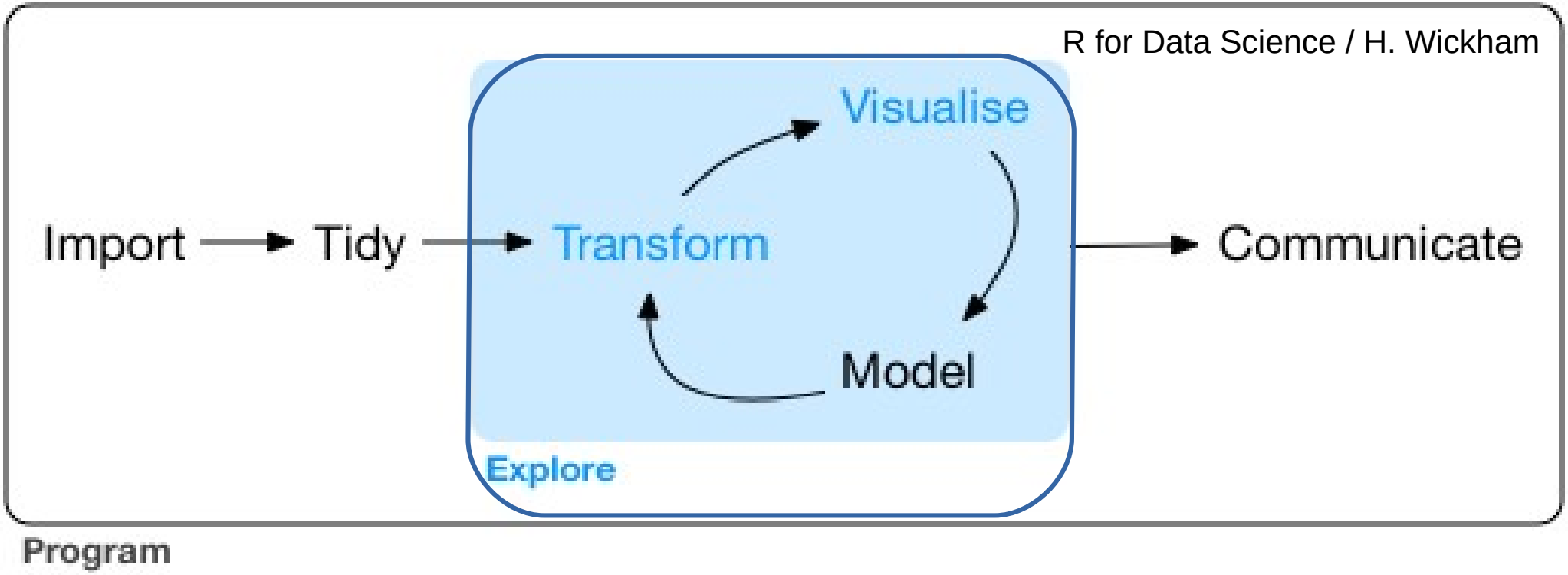
http://microbiome.github.io

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

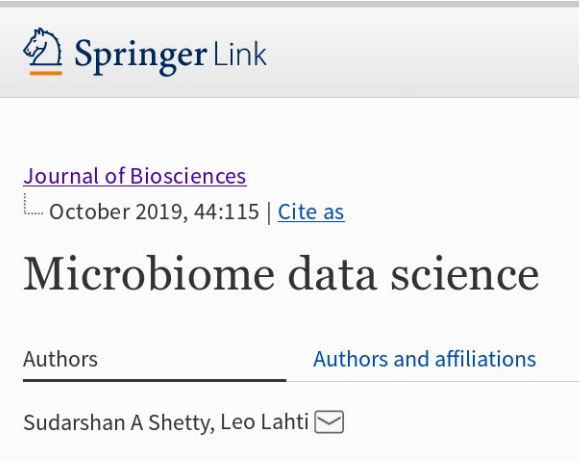
Microbiome data science

Sudarshan Shetty[1], Leo Lahti[2*]



Open Data Science

In: Leo Lahti. Advances in Intelligent Data Analysis XVII. Lecture Notes in Computer Science 11191., Springer Nature, India, 2018. Conference proceedings.



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Education

A Quick Guide to Software Licensing for the Scientist-Programmer

Andrew Morin[1], Jennifer Urban[2], Piotr Sliz[1]*

PeerJ Computer Science

Software citation principles

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

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

ABSTRACT

Software is a critical part of modern research and yet there is little support across the scholarly ecosystem for its acknowledgement and citation. Inspired by the activities

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

What is  ?

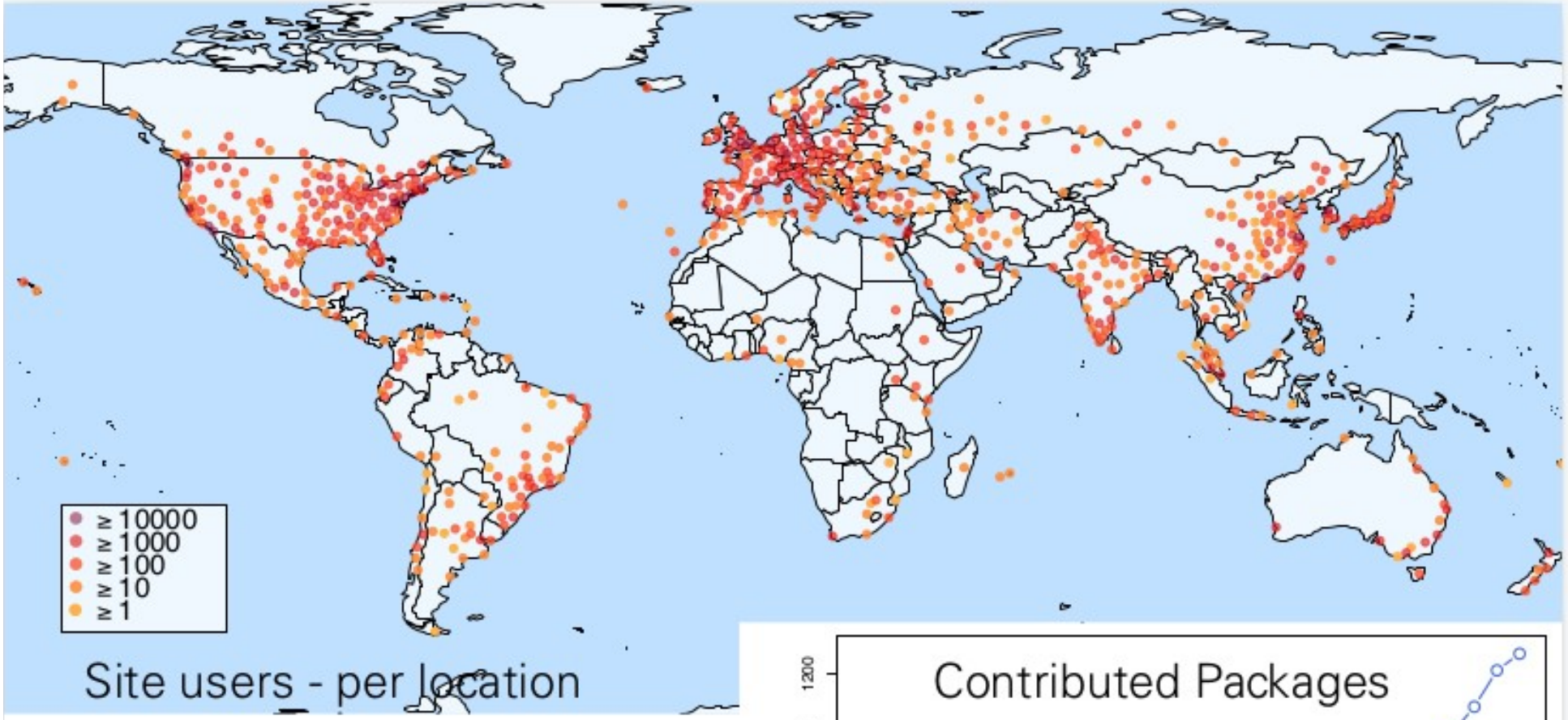
Started 2001 as a platform for analysis & understanding of microarray data

More than 1,600 packages. Domains of expertise:

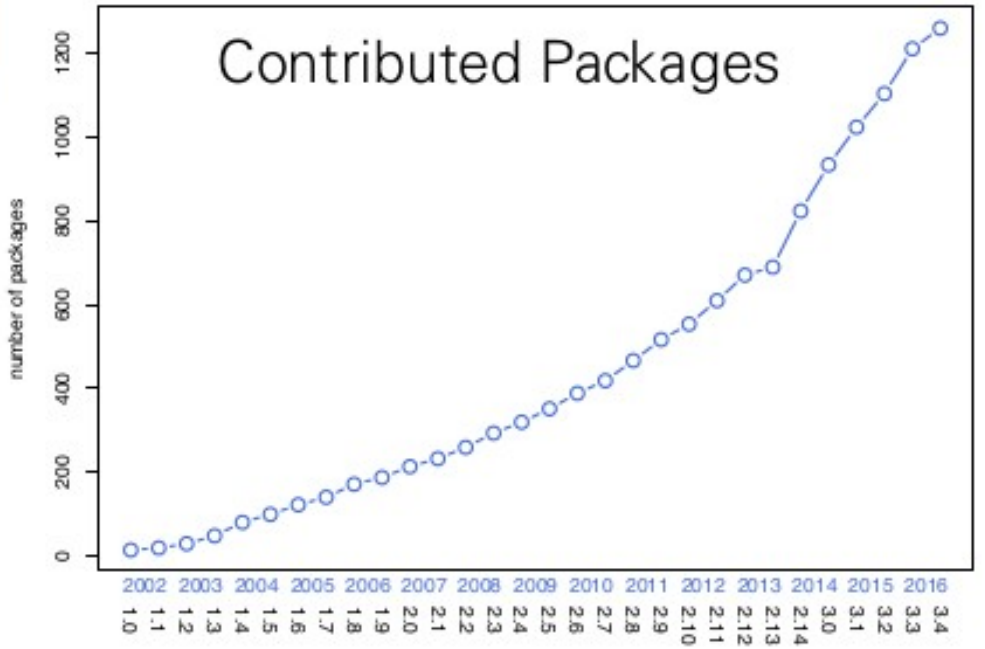
- Sequencing (RNASeq, ChIPSeq, single-cell, called variants, ...)
- Microarrays (methylation, expression, copy number, ...)
- Flow cytometry
- Proteomics
- Multi-Omics data integration

Important themes

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



Site users - per location



World largest bioinformatics project

10,000s users

>18,000 papers in PubmedCentral

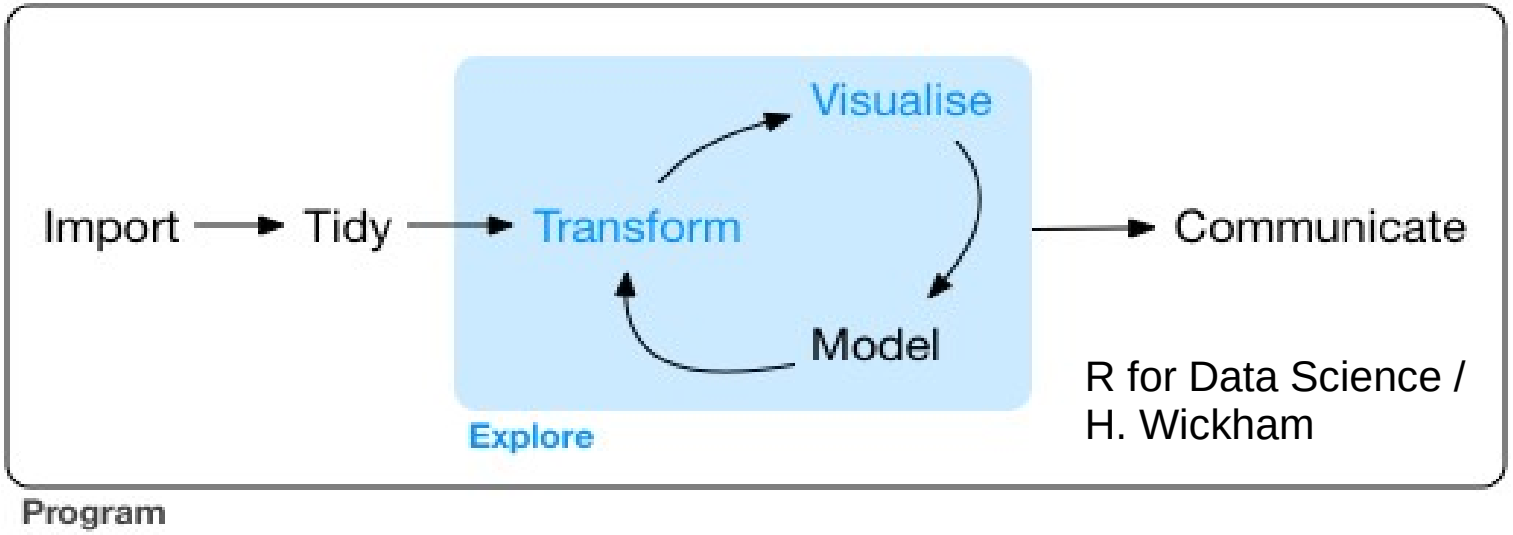
REVISED

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[1]

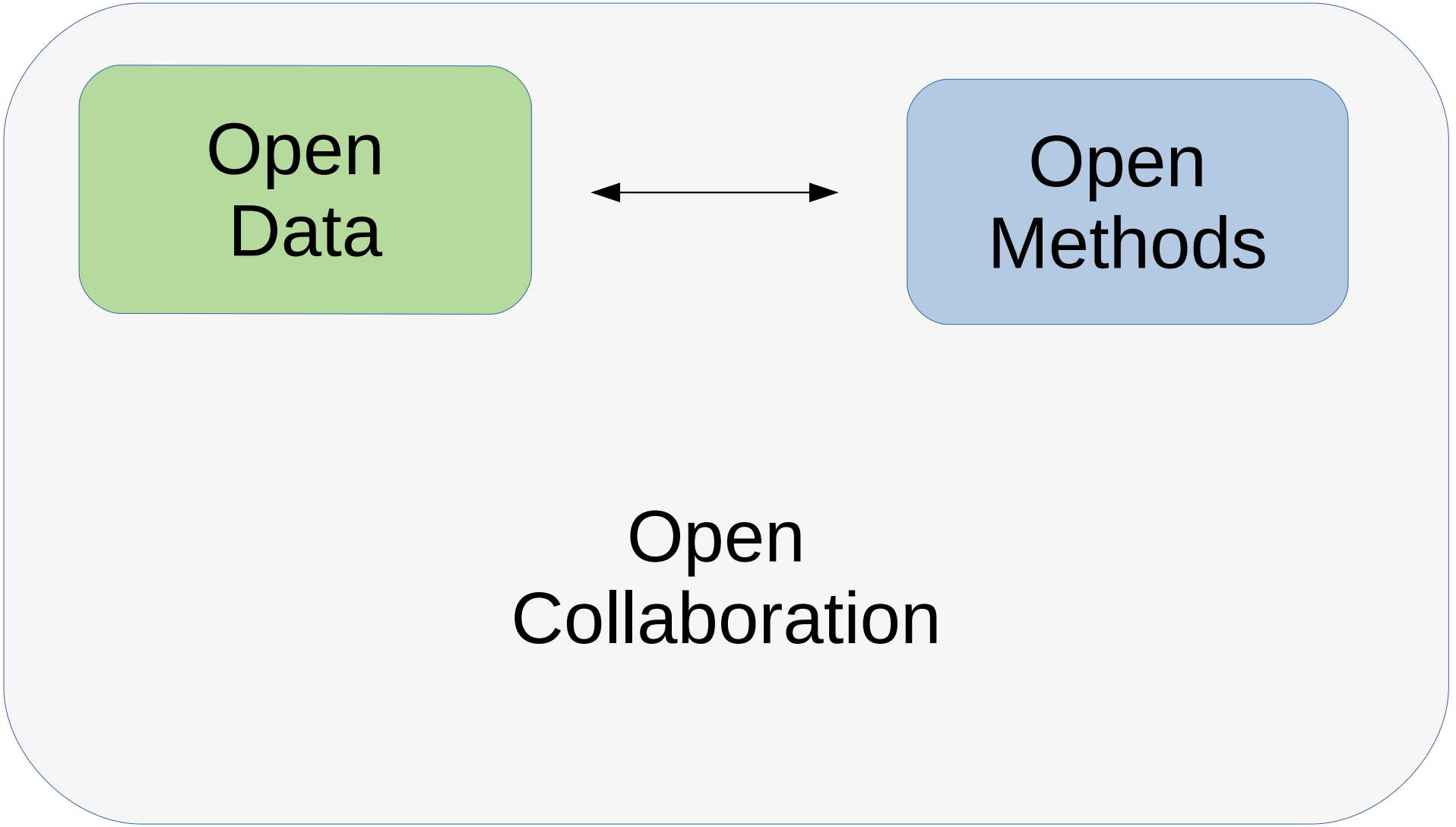
Author details

This article is included in the Bioconductor gateway.



Open data science

Leo Lahti. In: Advances in Intelligent Data Analysis XVII.
Lecture Notes in Computer Science 11191., Springer Nature, 2018.



NATURE | TOOLBOX

How scientists use Slack

Eight ways labs benefit from the popular workplace messaging tool.

Jeffrey M. Perkel

29 December 2016

PDF | Rights & Permissions

https://sdacrew.slack.com/

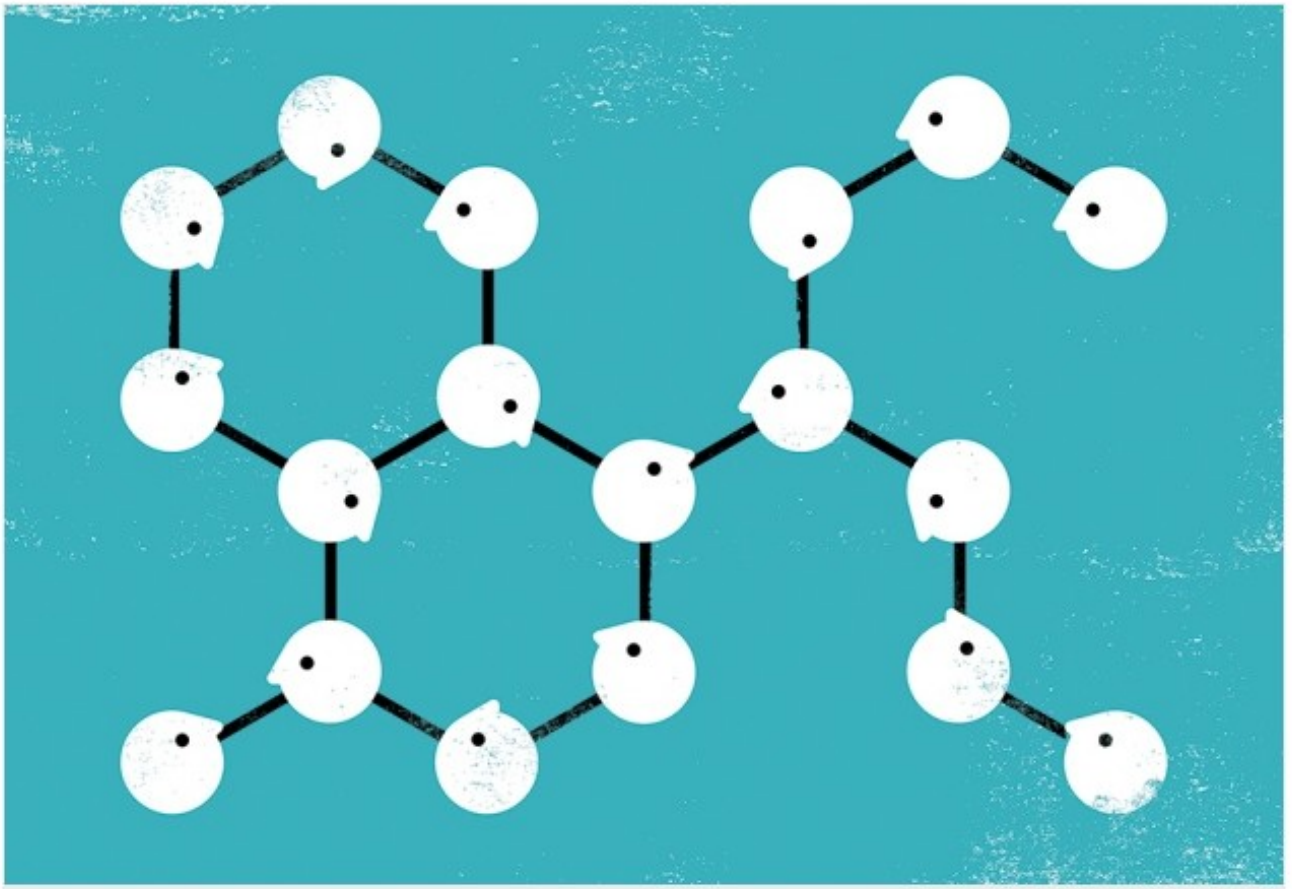
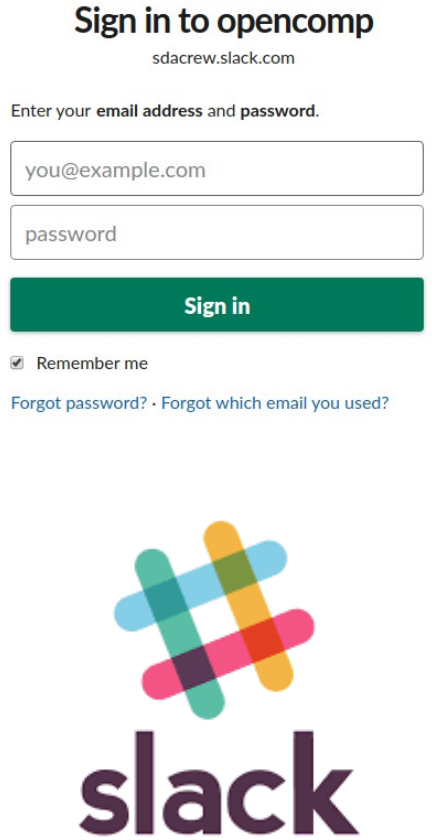


Illustration by the project twins



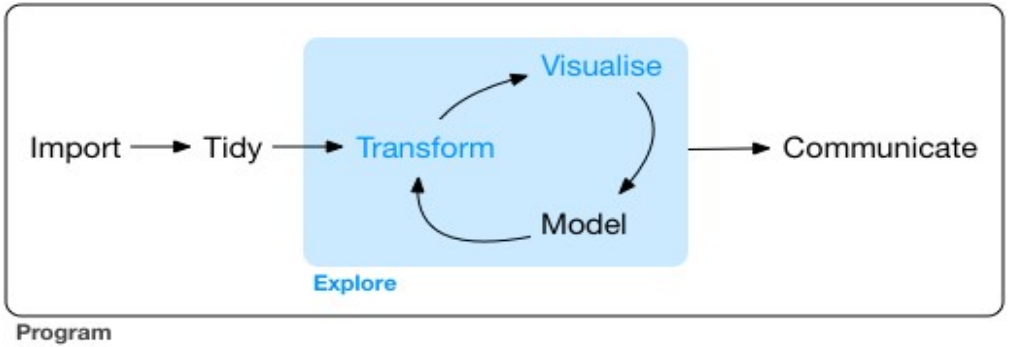
Labs

Data formats: Phyloseq

Data preprocessing: DADA2

Data manipulation: Toolkit

Reproducible research: Rmarkdown



Acknowledgments (slides, materials)

- Susan Holmes
- Wolfgang Huber
- Mike Lee
- Sudarshan Shetty
- Wisam Saleem