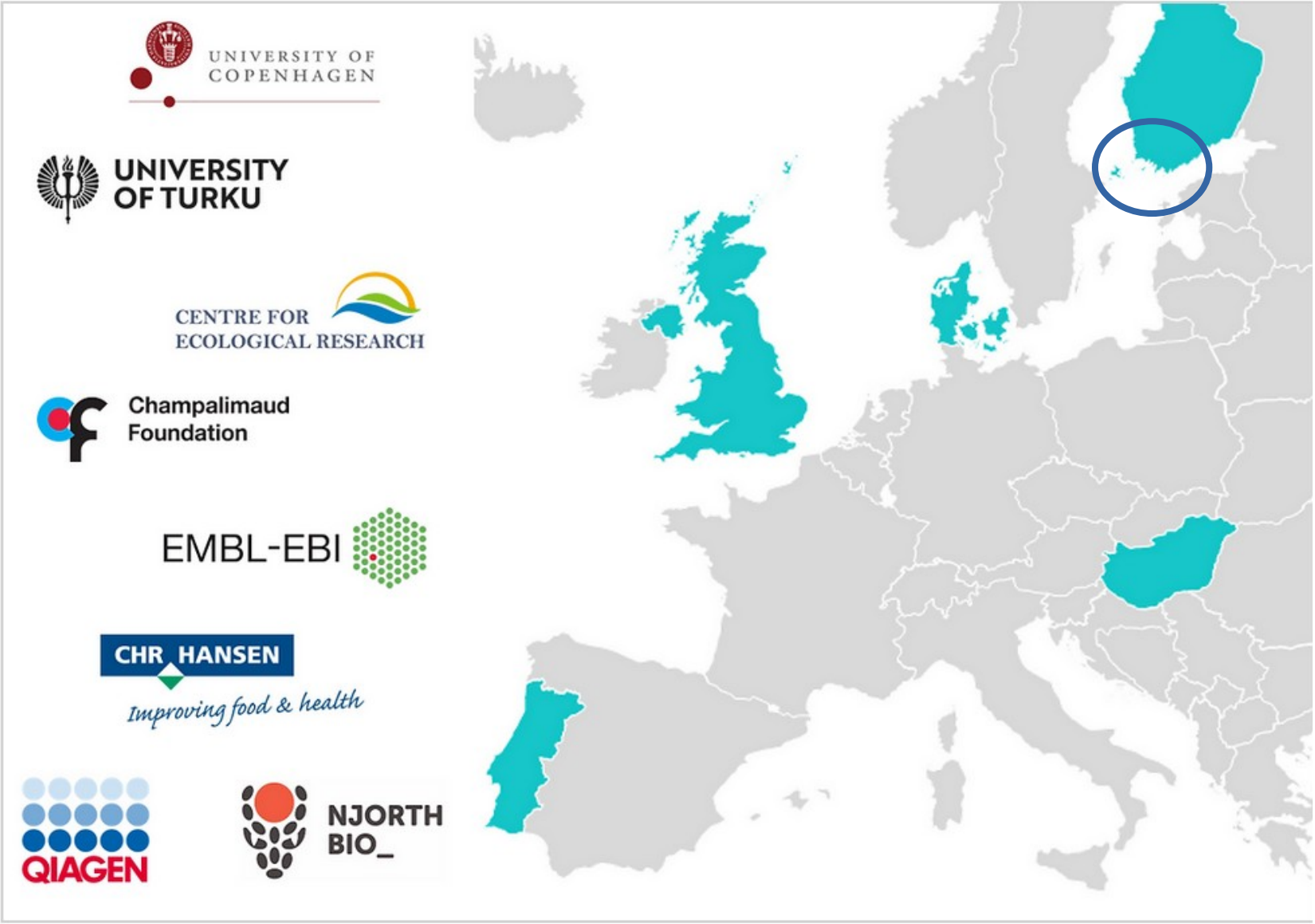
Introduction to multi-omics data analysis & machine learning workshop

H2020/FindingPheno, Jan 13-14, 2022

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

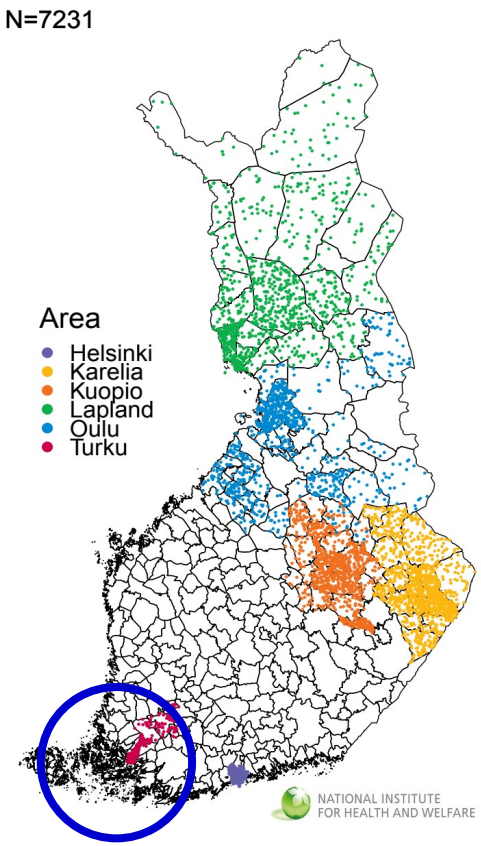
Turun yliopisto
University of Turku

FindingPheno is a truly interdisciplinary project featuring close collaboration between academic and industrial partners. Our combined scientific expertise and experiences include statistics and machine learning, ecology and evolutionary genetics, industrial food production systems experience and development of world class genomics software.



Department of Computing, University of Turku, Finland

datascience.utu.fi



The University of Turku

Turku, Finland

The University of Turku (UTU) was established in 1920 but its roots reach back to the Royal Academy of Turku in the 1650s. Today, UTU has almost 20,000 students and 3,271 staff members.

The UTU researchers participating in FindingPheno come from the Turku Data Science Group within the Department of Computing. This group is headed by Assoc. Prof. Leo Lahti and combines theory and methods of algorithmic data analysis, with a particular focus on probabilistic machine learning, complex systems, high-throughput data analysis and statistical programming.

FindingPheno

ABOUT | CONSORTIUM | BLOG | EVENTS | DELIVERABLES | CONTACT

Log In

Multi-omics data analysis for genotype-phenotype associations

The generation of large-scale biological data sets is exploding.
Genomics, transcriptomics, proteomics, metabolomics of both host and microbiome..
But! The methods to integrate and exploit this data do not keep up.

FindingPheno is answering this call.

Developing new computational tools for untangling host-microbiome interactions in multi-omic data sets

Find out more

EU-funded research collaboration with eight academic, research and industry partners from across Europe

Find out more

Validating our tools against commercially relevant test cases for more sustainable food production

Find out more

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952914

Contact us: findingpheno.eu@gmail.com

 Data and Accessibility Policy

Day 1 (Times in CET)

Lectures (45 min + 15 min breaks)

9:15-10:00 - **Welcome & introduction** - Leo Lahti, Associate professor (UTU)

10:15-11:00 - **Metagenomics** - Katariina Pärnänen, Postdoctoral researcher (UTU)

11:15-12:00 - **Metabolomics** - Pande Putu Erawijantari, Postdoctoral researcher (UTU)

12:15-13:00 - **Multiomics** - Leo Lahti, Associate professor (UTU)

13:00-14 - **Lunch** break

Practical session

14:15-17:00 - Tuomas Borman and Chouaib Benchraka, Research assistants (UTU)

Data import and data structures

Microbiome data exploration & visualization

Day 2 (Times in CET)

Lectures

9:15-10:00 - **Unsupervised ML**- Matti Ruuskanen, Postdoctoral researcher (UTU)

10:15-11:00 - **Supervised ML** - Matti Ruuskanen

11:15-12:00 - **Individual-based modeling** - Gergely Boza, Research fellow (CER)

12:15-13:00 - **Data integration** - Leo Lahti, Associate professor (UTU)

13:00-14 - **Lunch** break

Practical

14:15-17:00 - Tuomas Borman, Matti Ruuskanen and Chouaib Benchraka (UTU)

Unsupervised learning: Beta-diversity and biclustering

Supervised learning: Regression and classification with random forests

Validation and interpretation of black box models

Matti Ruuskanen
Postdoc
Turku, Finland

Pande Putu Erawijantari
Postdoc
Turku, Finland

Katariina Pärnänen
Postdoc
Turku, Finland

Gergely Boza
Research fellow,
Center for Ecological Research,
Hungary

Chouaib Benchraka
Scientific programmer
Turku, Finland

Tuomas Borman
Scientific programmer
Turku, Finland

Leo Lahti
Associate Prof.
Turku, Finland

Learning objectives

motivation & challenges in multi-omics

examples of computational approaches

hands-on experience on R/Bioconductor tools

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.*

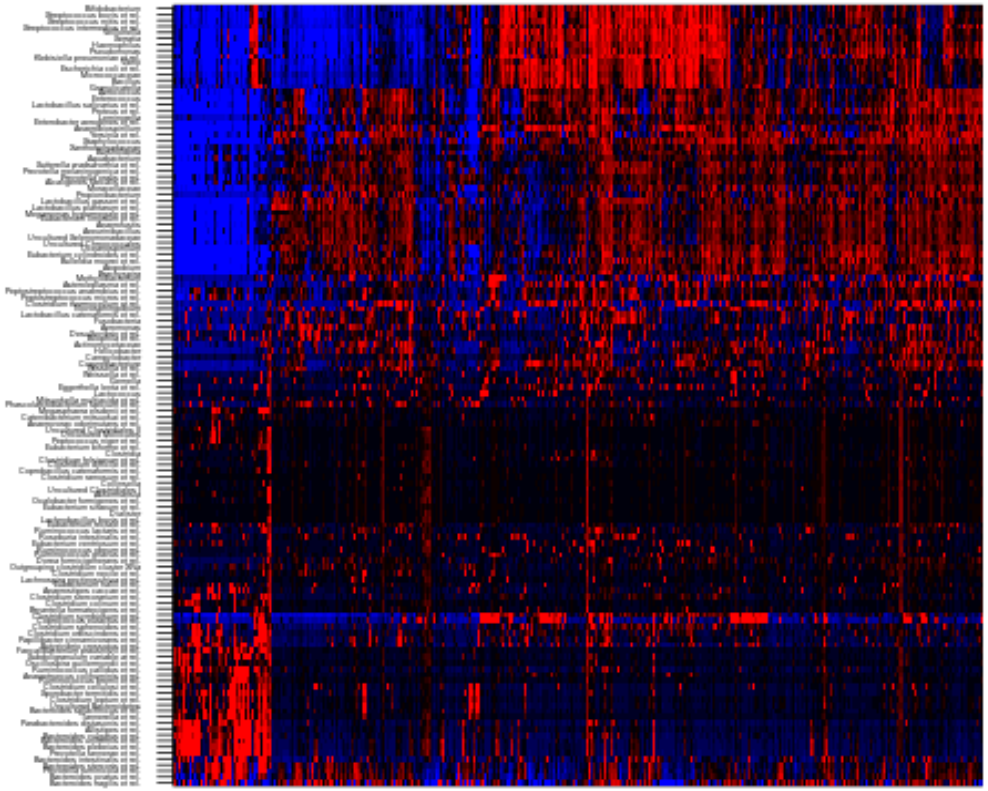
The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. **Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome**. We also present an initial analysis of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

Omics:
taxonomic abundance table

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

Individuals

Taxonomic groups



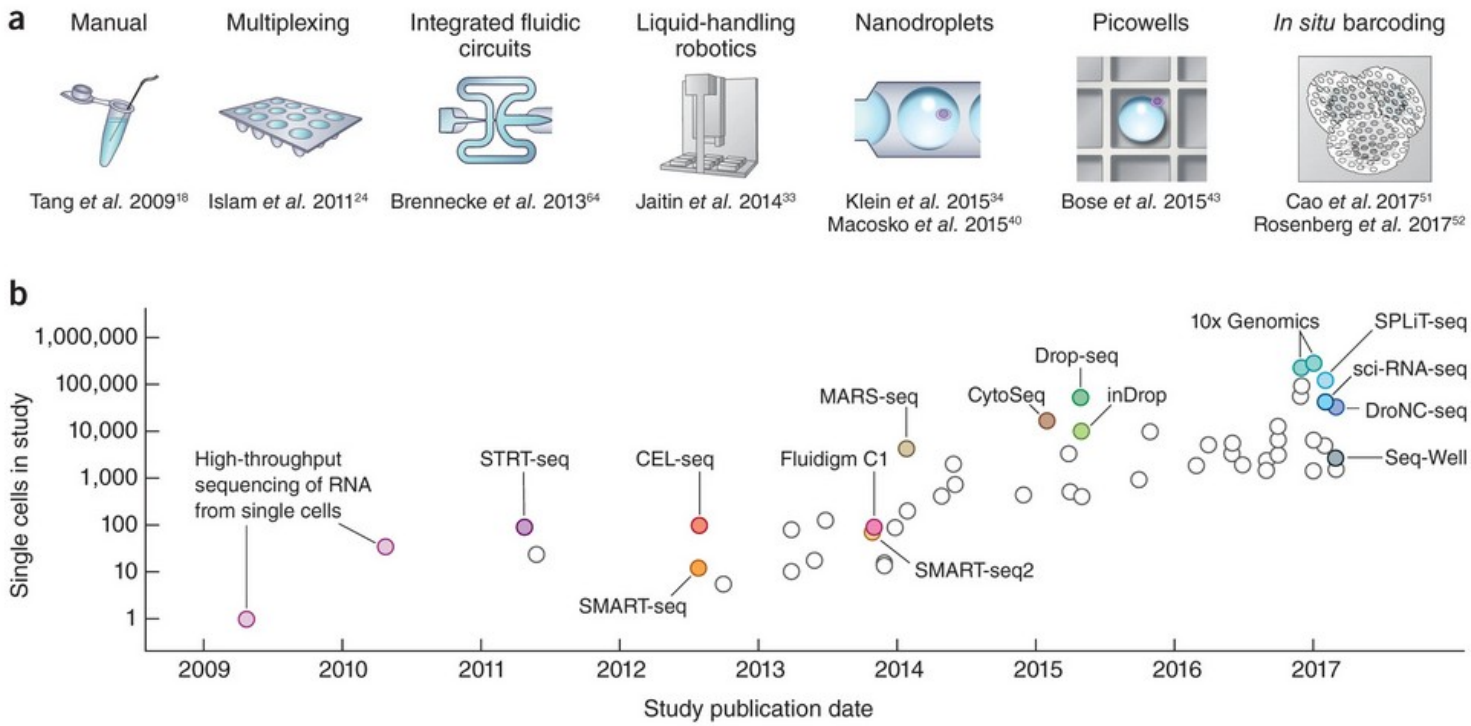
Genomics
Epigenomics
Microbiomics
Lipidomics
Proteomics
Glycomics
Foodomics
Transcriptomics
Metabolomics
Culturomics

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

Figure 1: Scaling of scRNA-seq experiments.

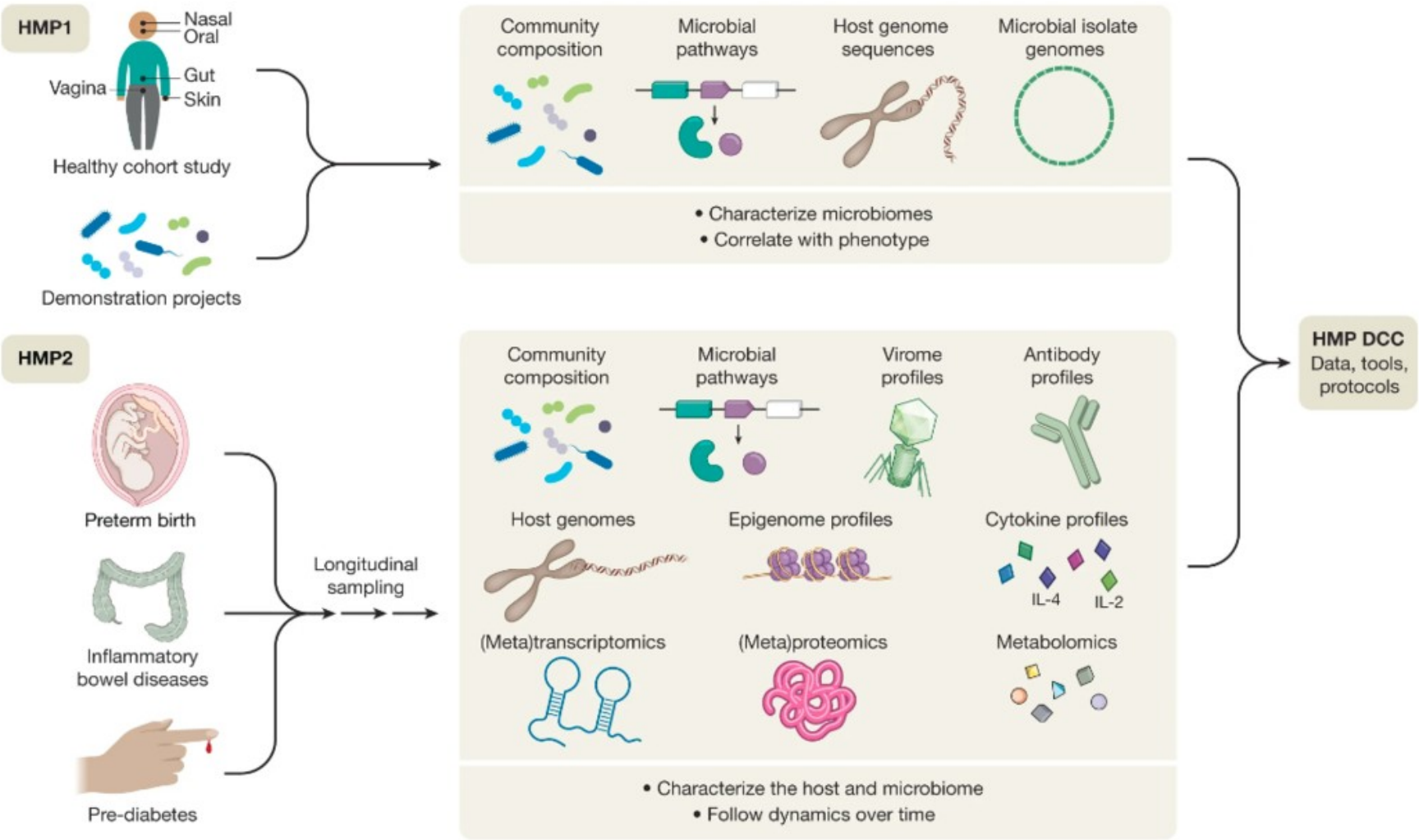
:he past

From: Exponential scaling of single-cell RNA-seq in the past decade



a) Key technologies that have allowed jumps in experimental scale. A jump to ∼100 cells was enabled by sample multiplexing, and then a jump to ∼1,000 cells was achieved by large-scale studies using integrated fluidic circuits, followed by a jump to several thousands of cells with liquid-handling robotics. Further orders-of-magnitude increases bringing the number of cells assayed into the tens of thousands were enabled by random capture technologies using nanodroplets and picowell technologies. Recent studies have used in situ barcoding to inexpensively reach the next order of magnitude of hundreds of thousands of cells. (b) Cell numbers reported in representative publications by publication date. Key technologies are indicated.

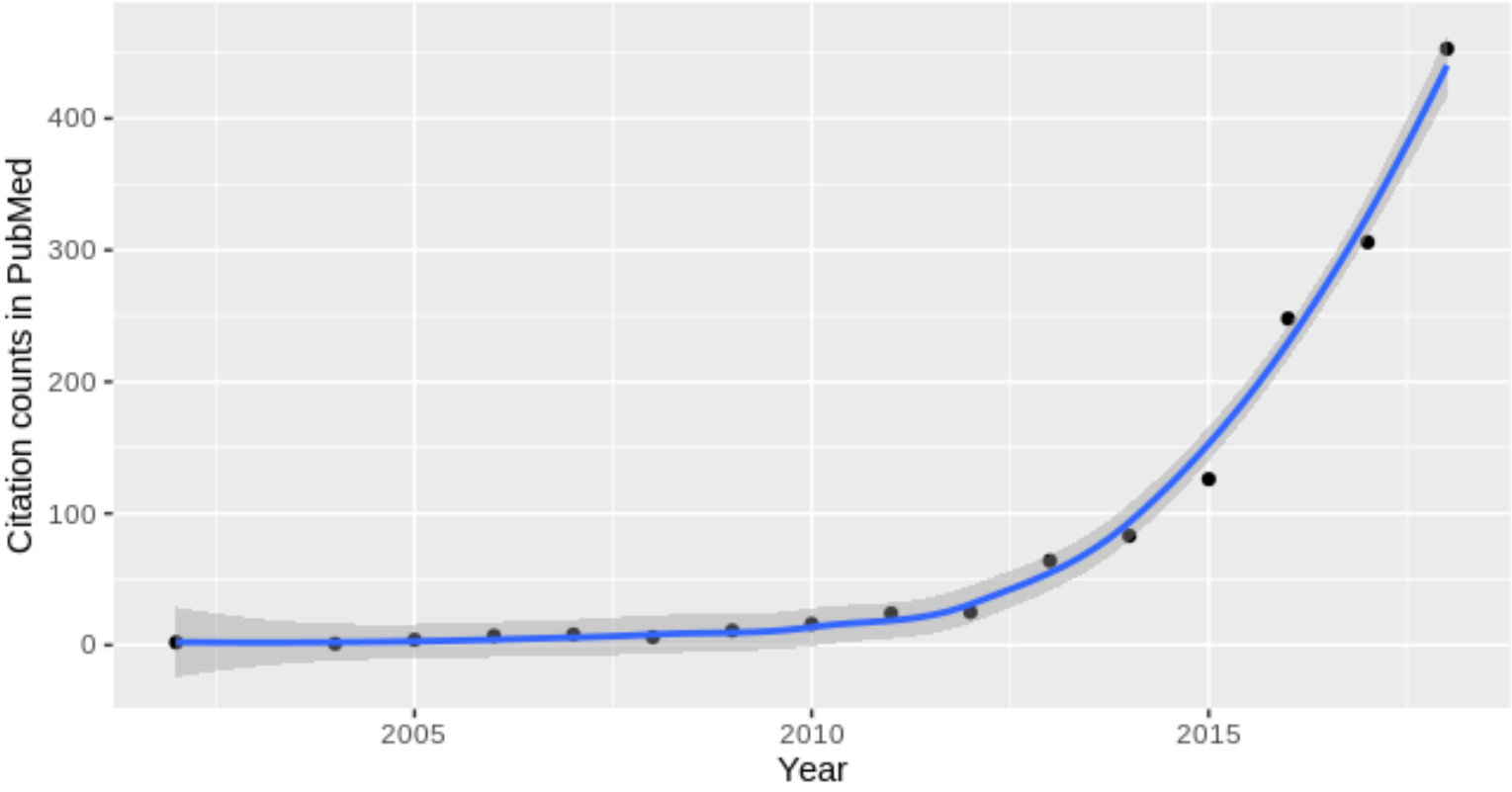
Multi-omics



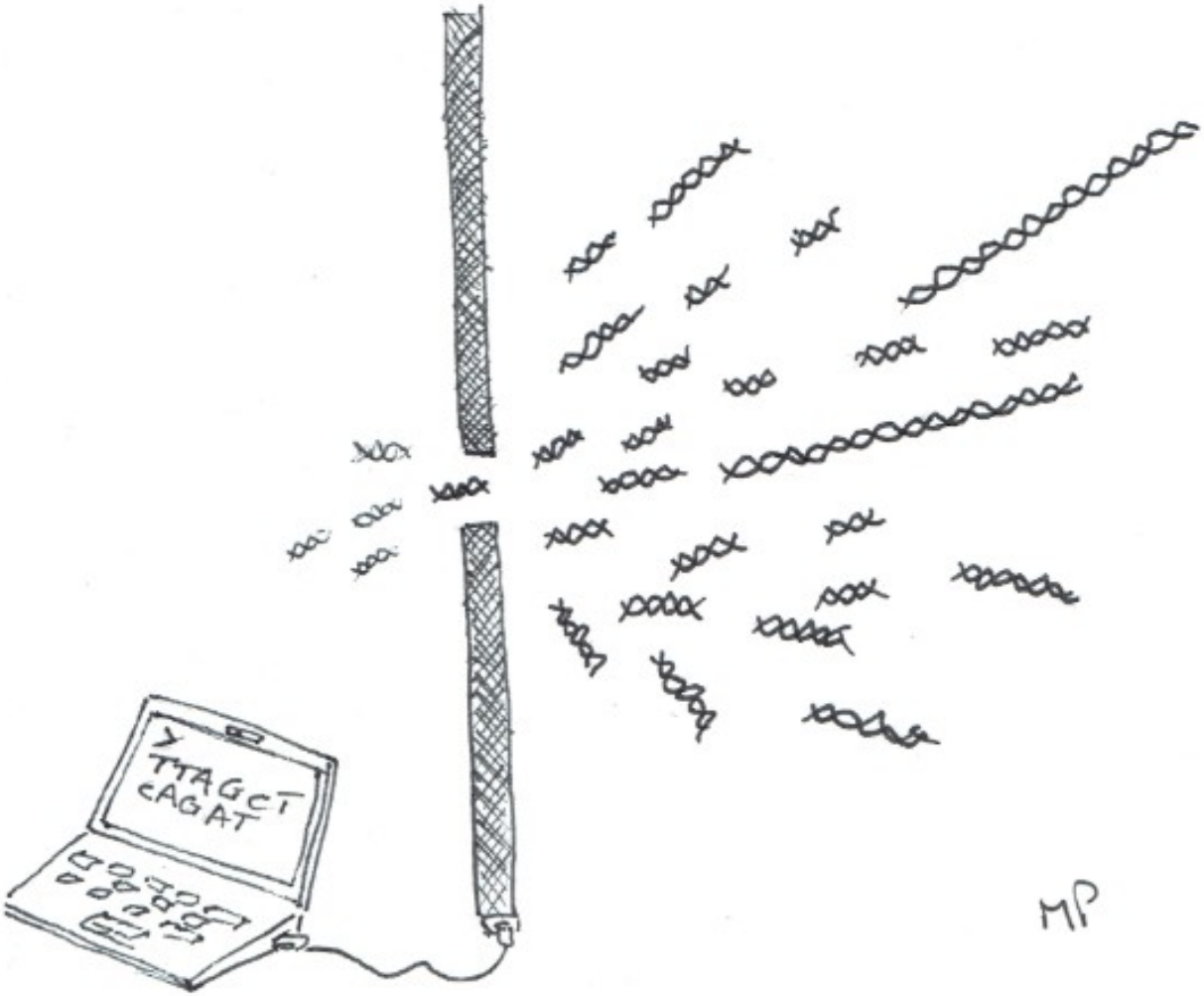
Overview of phases 1 and 2 of the human microbiome project. Source: Wikipedia.

Number of articles related to "multiomics" in PubMed until 2018

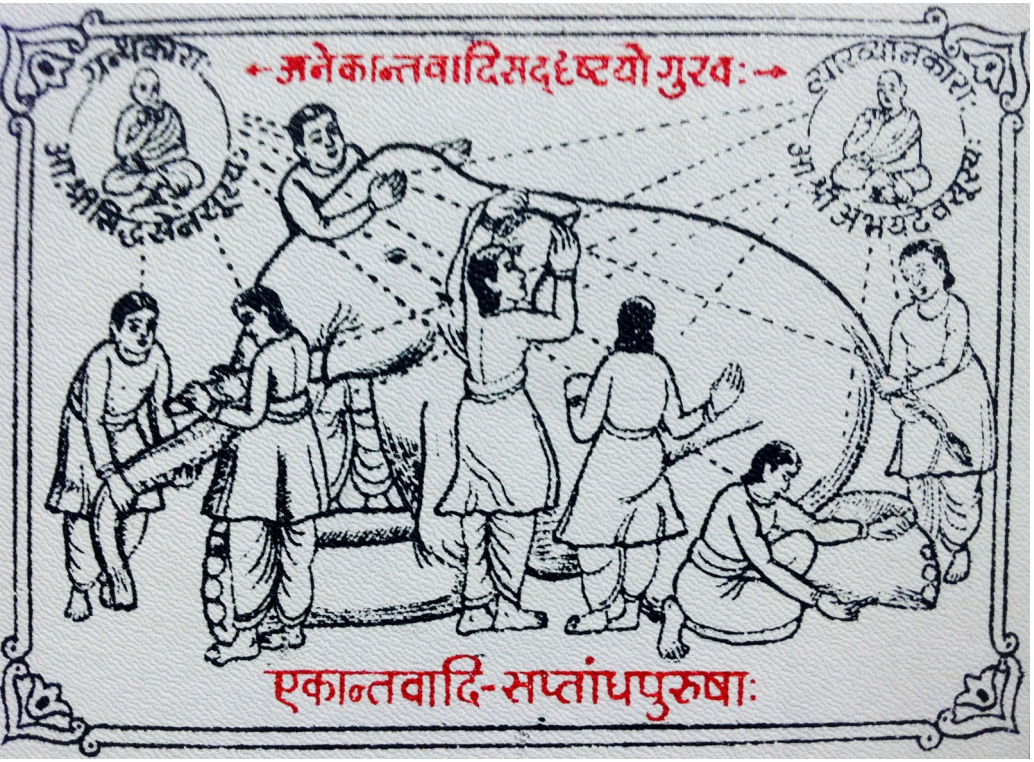
(source: Wikipedia)



Limited observations → data integration?



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

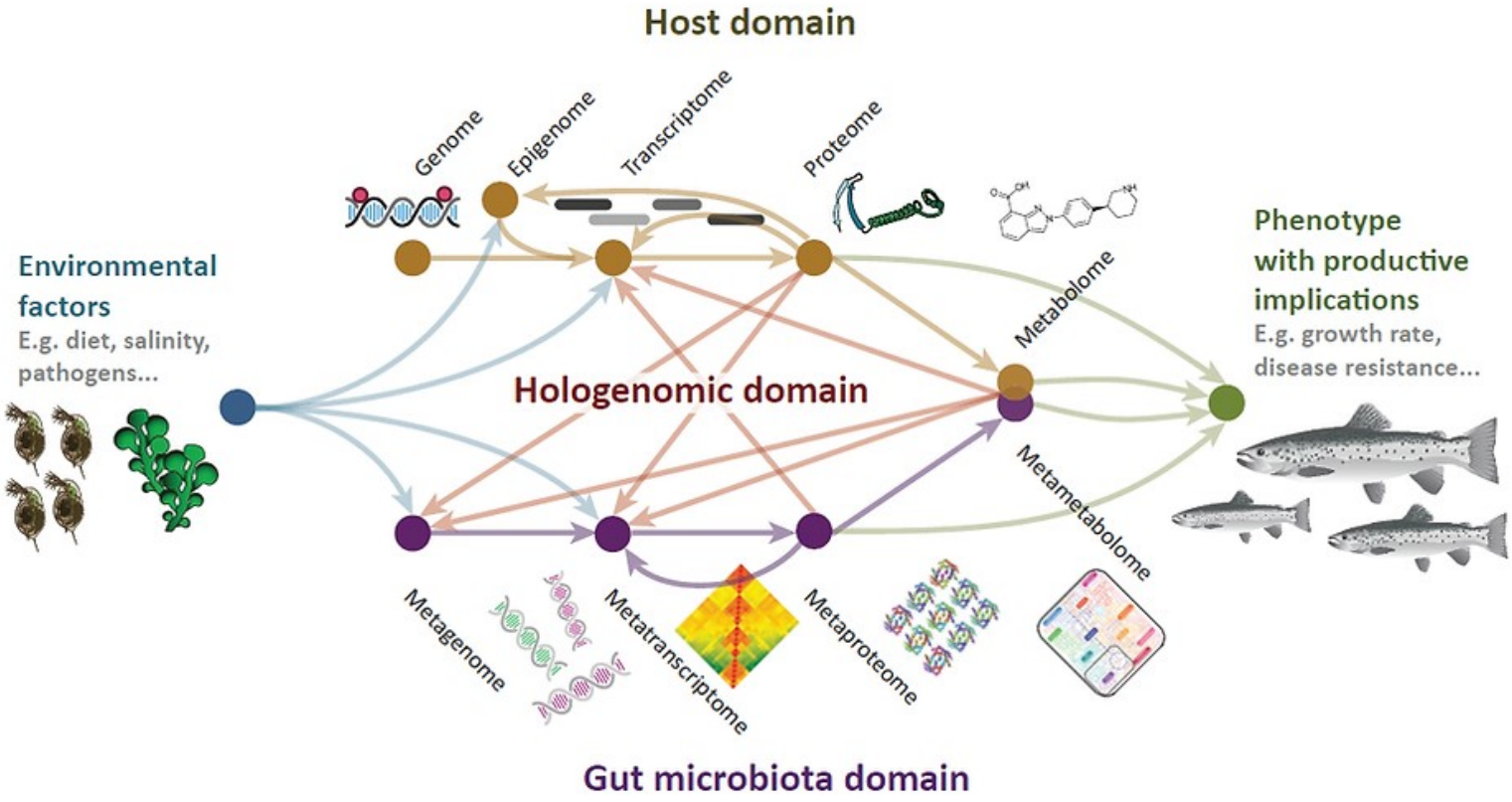


Elephant in the dark

The medieval era Jain texts explain the concepts of anekāntavāda (or "many-sidedness") and syādvāda ("conditioned viewpoints") with the parable of the blind men and an elephant (Andhgajanyāyah), which addresses the manifold nature of truth.

The Buddhist text Tittha sutta, Udāna 6.4, Khuddaka Nikaya, contains one of the earliest versions of the story. The Tittha sutta is dated to around c. 500 BCE, although the parable is likely older.

FindingPheno is creating an integrated computational framework for hologenomic big data, providing the tools to better understand how host-microbiome interactions can affect growth and other outcomes.



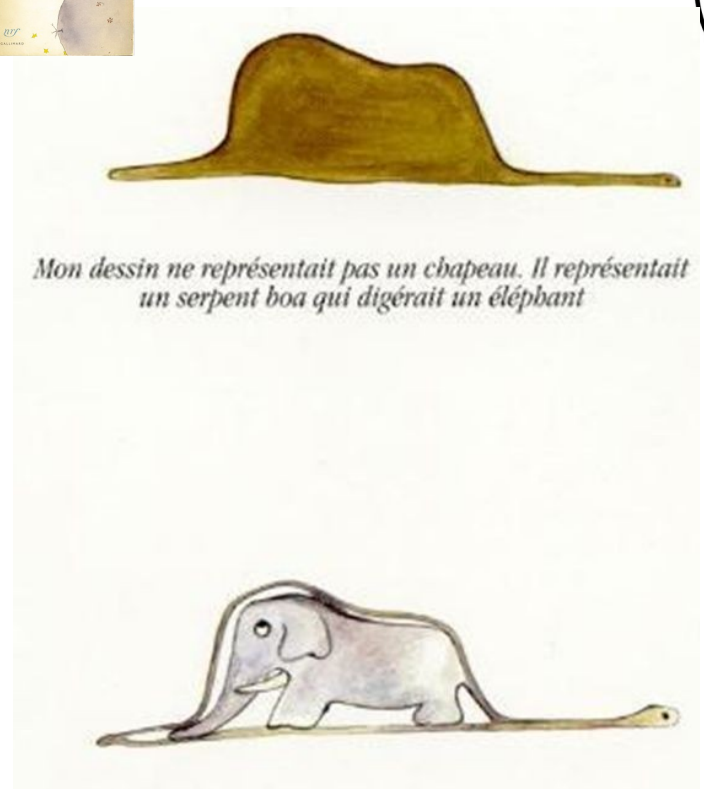
Understanding the hologenomic domain is a fiendishly difficult problem, with a complex tangle of interactions at many molecular levels both within and between organisms. FindingPheno aims to solve this problem, developing a unified statistical framework for the intelligent integration of multi-omic data from both host and microbiome to understand biological outcomes.

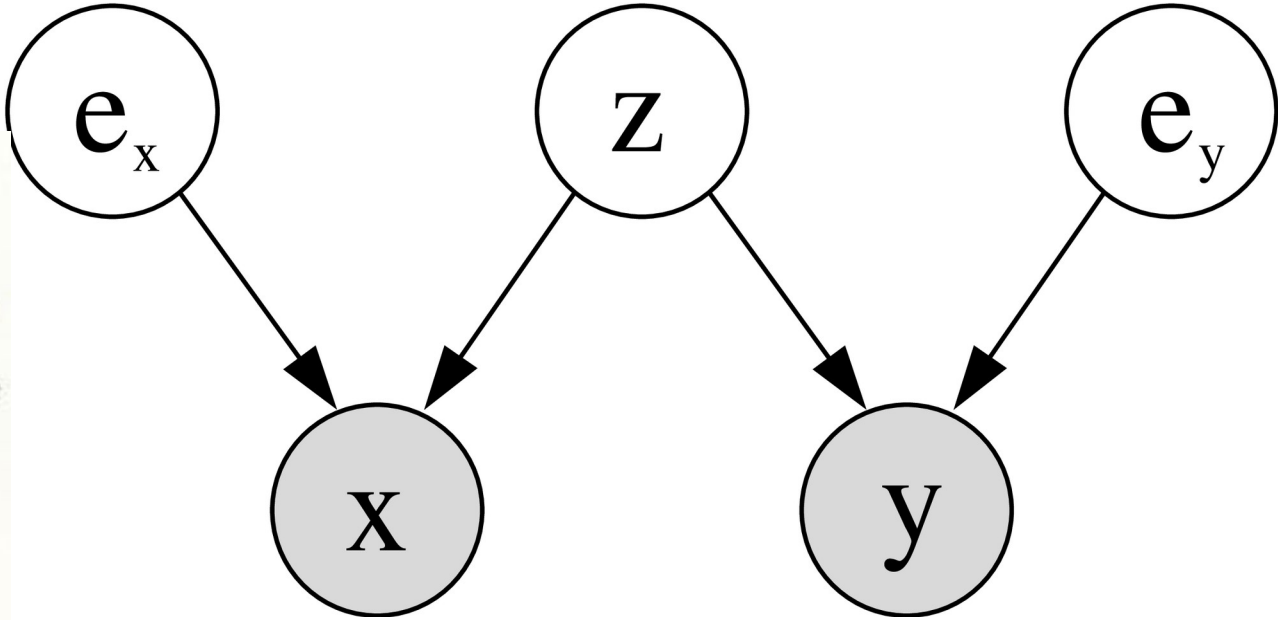
We apply state-of-the-art mathematical and machine learning approaches taken from evolutionary genomics, collective behaviour analysis, ecosystem dynamics, statistical modelling, and applied agricultural research to give us a truly interdisciplinary perspective towards solving this difficult problem. Our project takes a unique two-pronged approach: combining biology-agnostic machine learning methods with biology-informed hierarchical modelling to increase the power and adaptability of our predictive tools.

The tools created in FindingPheno are expected to significantly improve how we understand and utilise the functions provided by microbiomes in combating human diseases as well as the way we produce sustainable food for future generations.

$$\begin{cases} X = W_x \mathbf{z} + \varepsilon_x \\ Y = W_y \mathbf{z} + \varepsilon_y \end{cases}$$

Multi-view learning





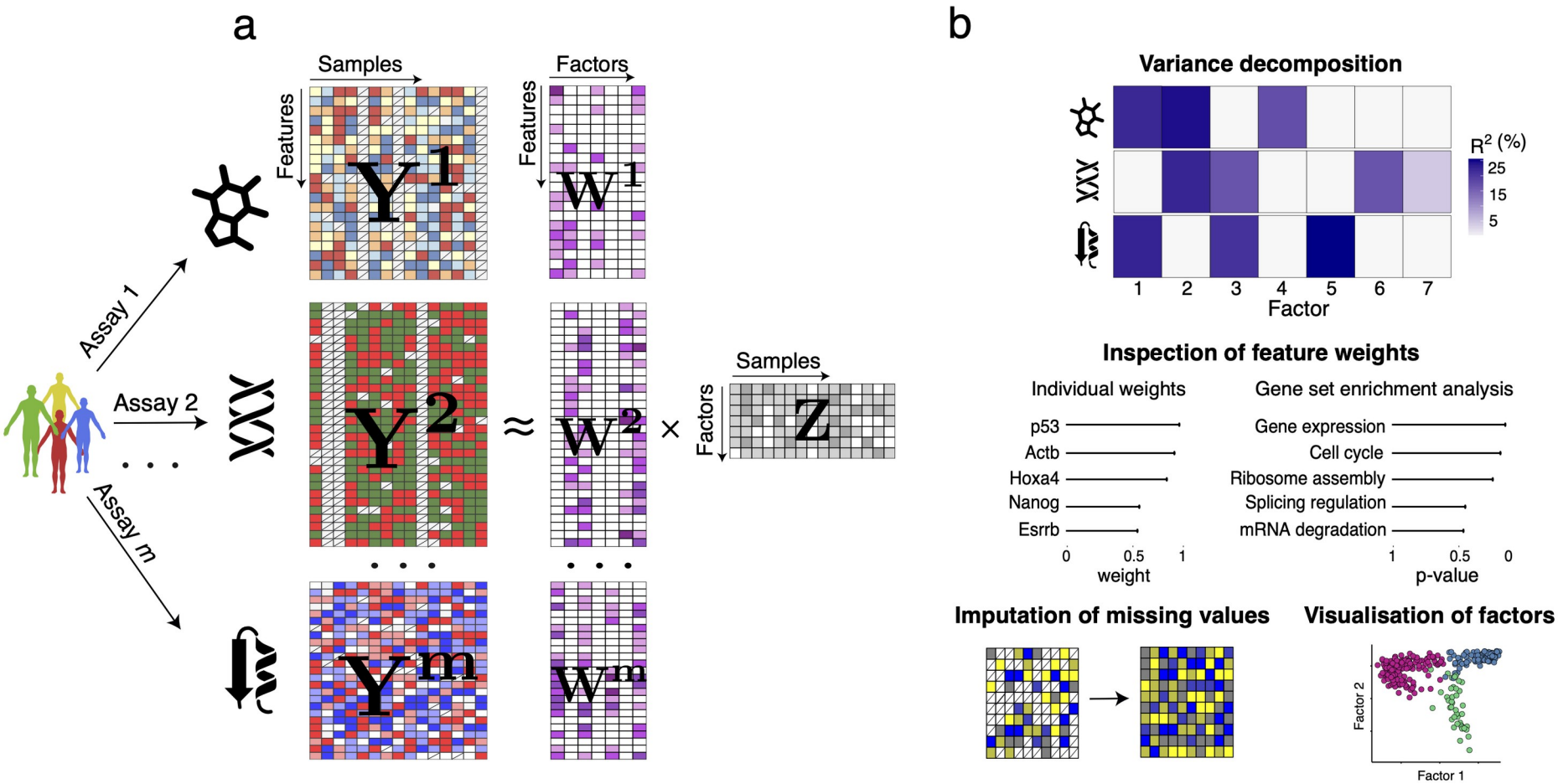
Method | 20 June 2018 | OPEN ACCESS TRANSPARENT PROCESS

Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets

Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, Oliver Stegle

Author Information

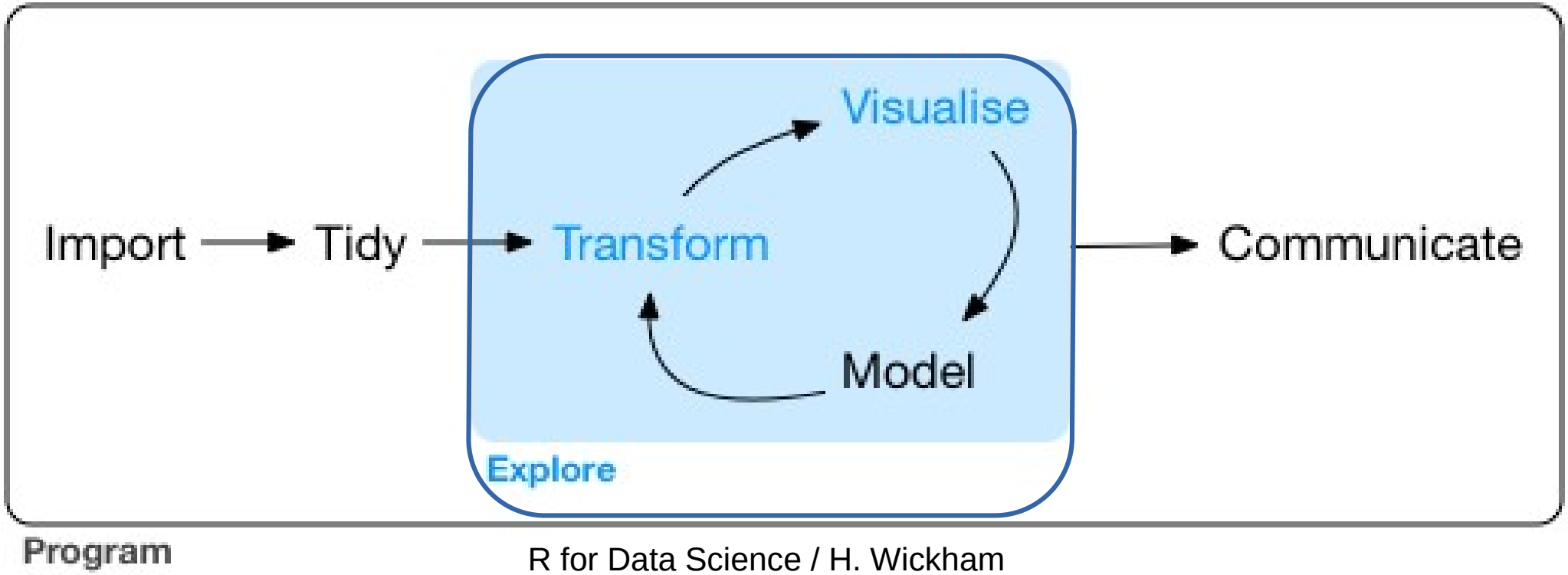
Molecular Systems Biology (2018) 14: e8124 | https://doi.org/10.15252/msb.20178124



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

open data science frameworks

Computational workflows have an increasingly central role in research!



R for Data Science / H. Wickham

The influence of hidden researcher decisions in applied microeconomics

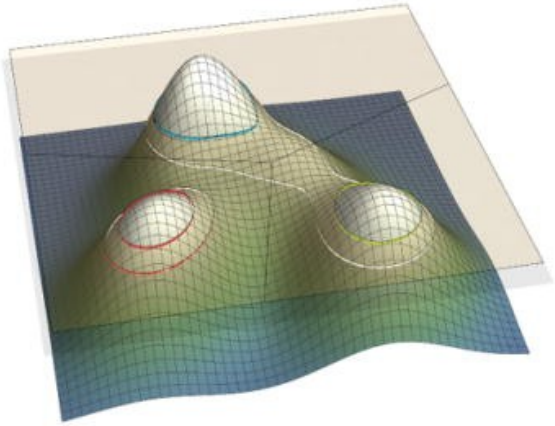
Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, Yaniv Stopnitzky

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.

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

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



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

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

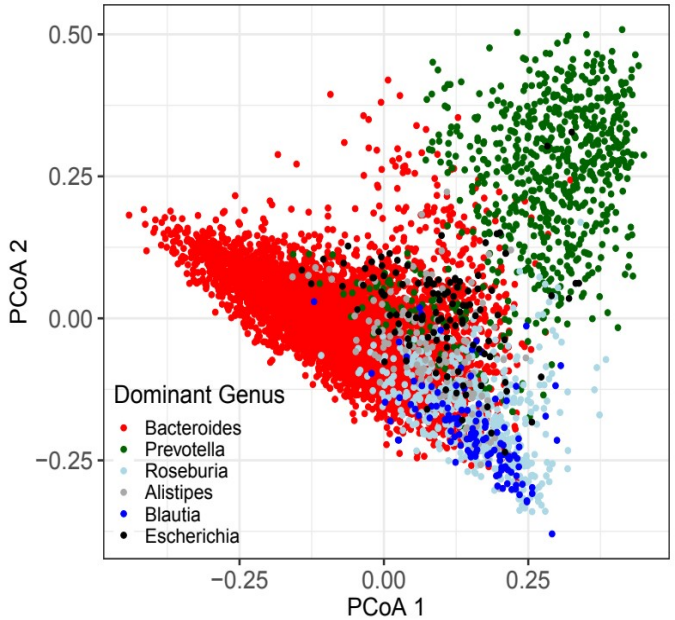
(Dis)similarity
- Eulidean
- Aitchison
- Bray-Curtis
- Jaccard
- weighted Unifrac
- unweighted Unifrac

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

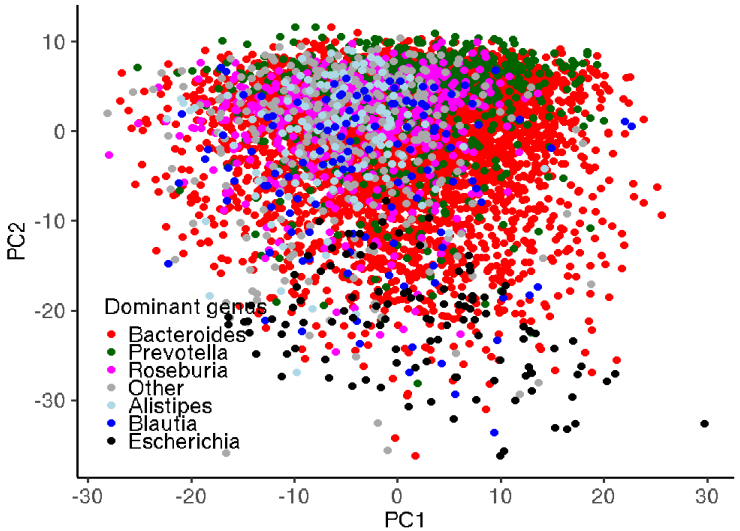
Regulation
- Calinski-Harabasz
- Dirichlet Process
- Silhouette Index
- AIC
- BIC
- DIC

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

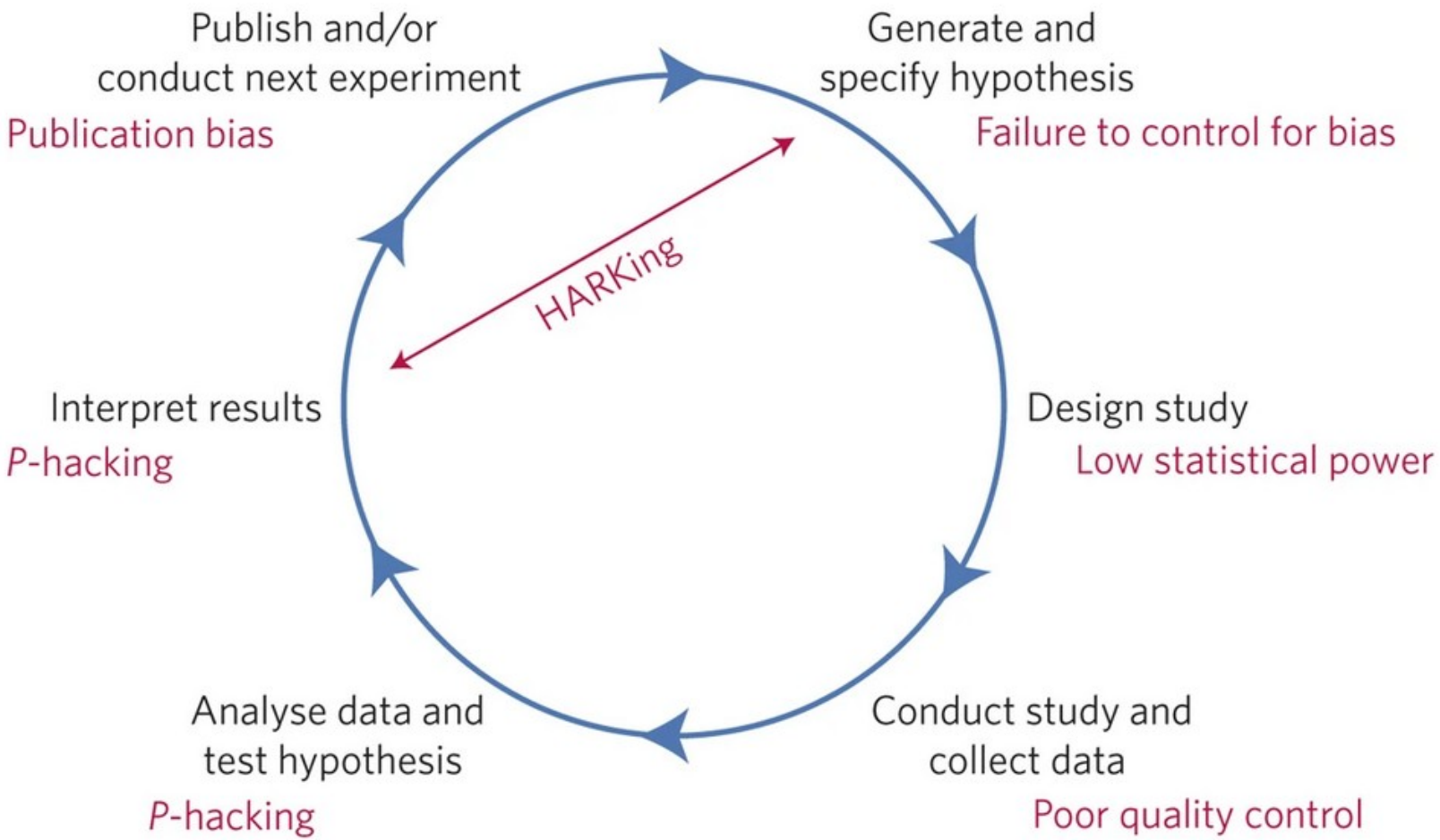
A manifesto for reproducible science

Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis

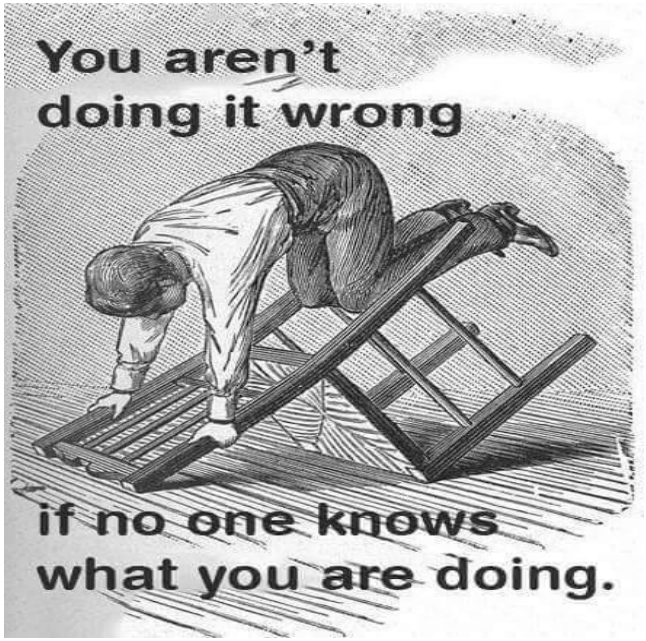
Nature Human Behaviour **1**, Article number: 0021 (2017) | Cite this article

Figure 1: Threats to reproducible science.

From: A manifesto for reproducible science



An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.



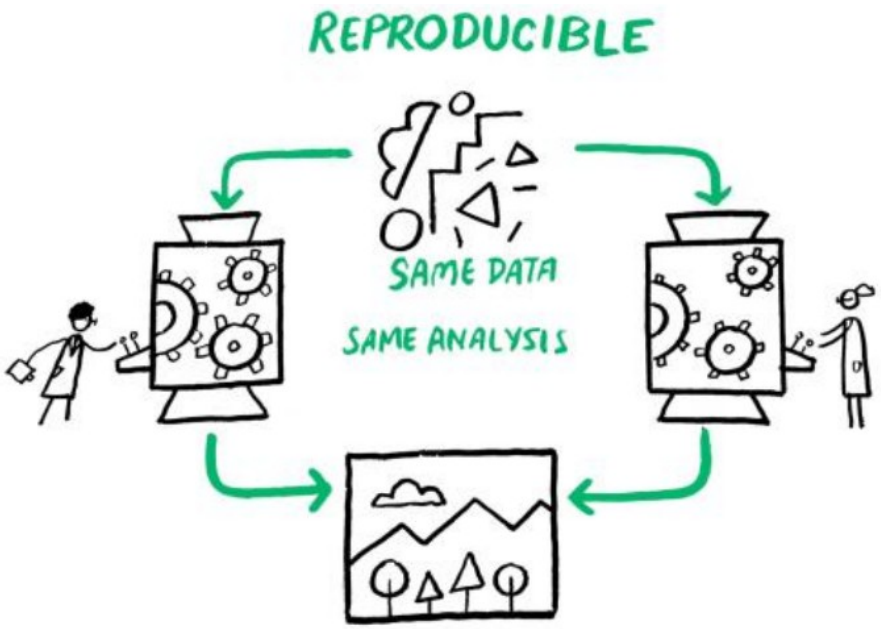
RESEARCH PRIORITIES

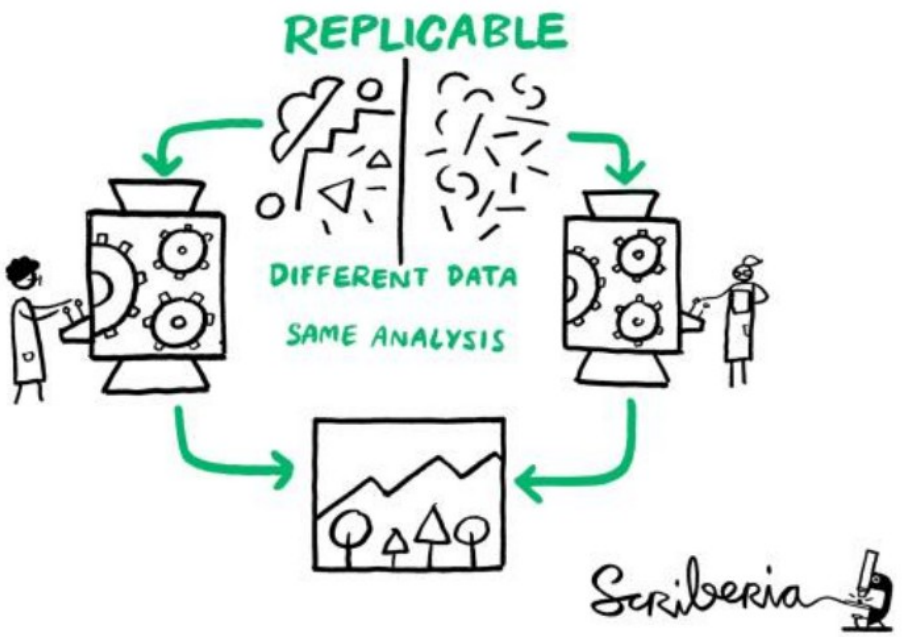
Shining Light into Black Boxes

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

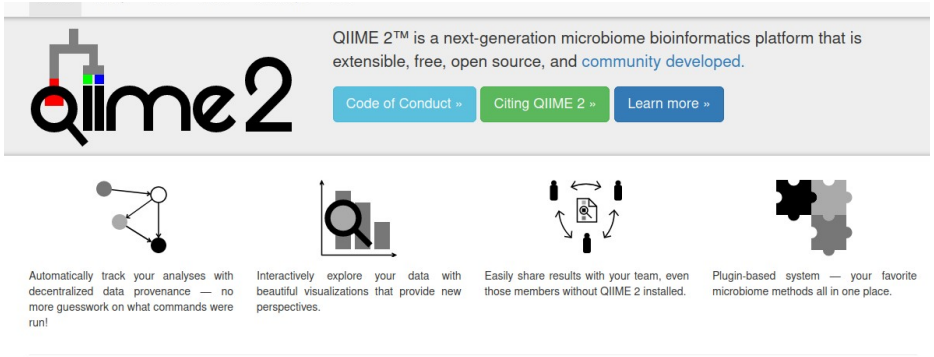
"*I have begun to think that no one ought to publish biometric results, without lodging a well arranged and well bound manuscript copy of all his data, in some place whereit should be accessible, under reasonable restrictions, to those who desire to verify his work.*"

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





open microbiome analysis frameworks



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

PeerJ ›

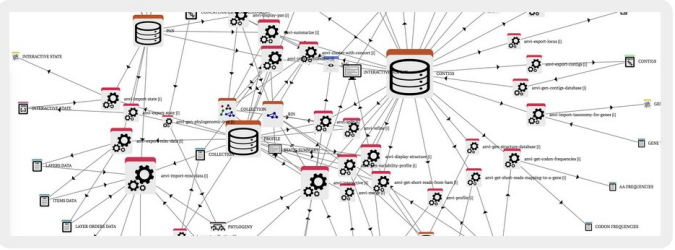
Anvi'o: an advanced analysis and visualization platform for 'omics data

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]

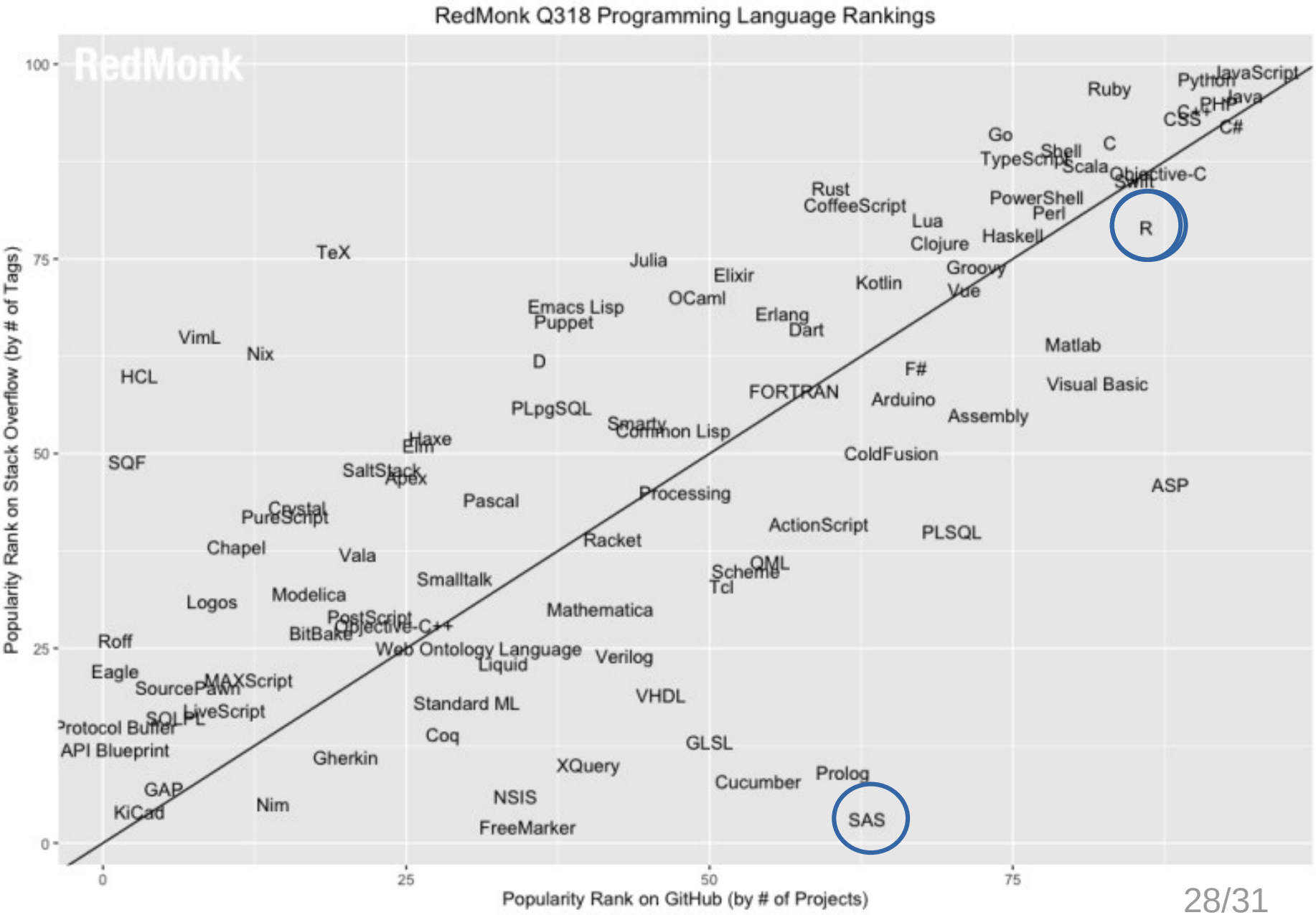
blished October 8, 2015

Anvi'o in a nutshell



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

Cultures of open data science collaboration



Day 1 (Times in CET)

Lectures (45 min + 15 min breaks)

9:15-10:00 - **Welcome & introduction** - Leo Lahti, Associate professor (UTU)

10:15-11:00 - **Metagenomics** - Katariina Pärnänen, Postdoctoral researcher (UTU)

11:15-12:00 - **Metabolomics** - Pande Putu Erawijantari, Postdoctoral researcher (UTU)

12:15-13:00 - **Multiomics** - Leo Lahti, Associate professor (UTU)

13:00-14 - **Lunch** break

Practical session

14:15-17:00 - Tuomas Borman and Chouaib Benchraka, Research assistants (UTU)

Data import and data structures

Microbiome data exploration & visualization

https://microbiome.github.io/course_2022_FindingPheno/

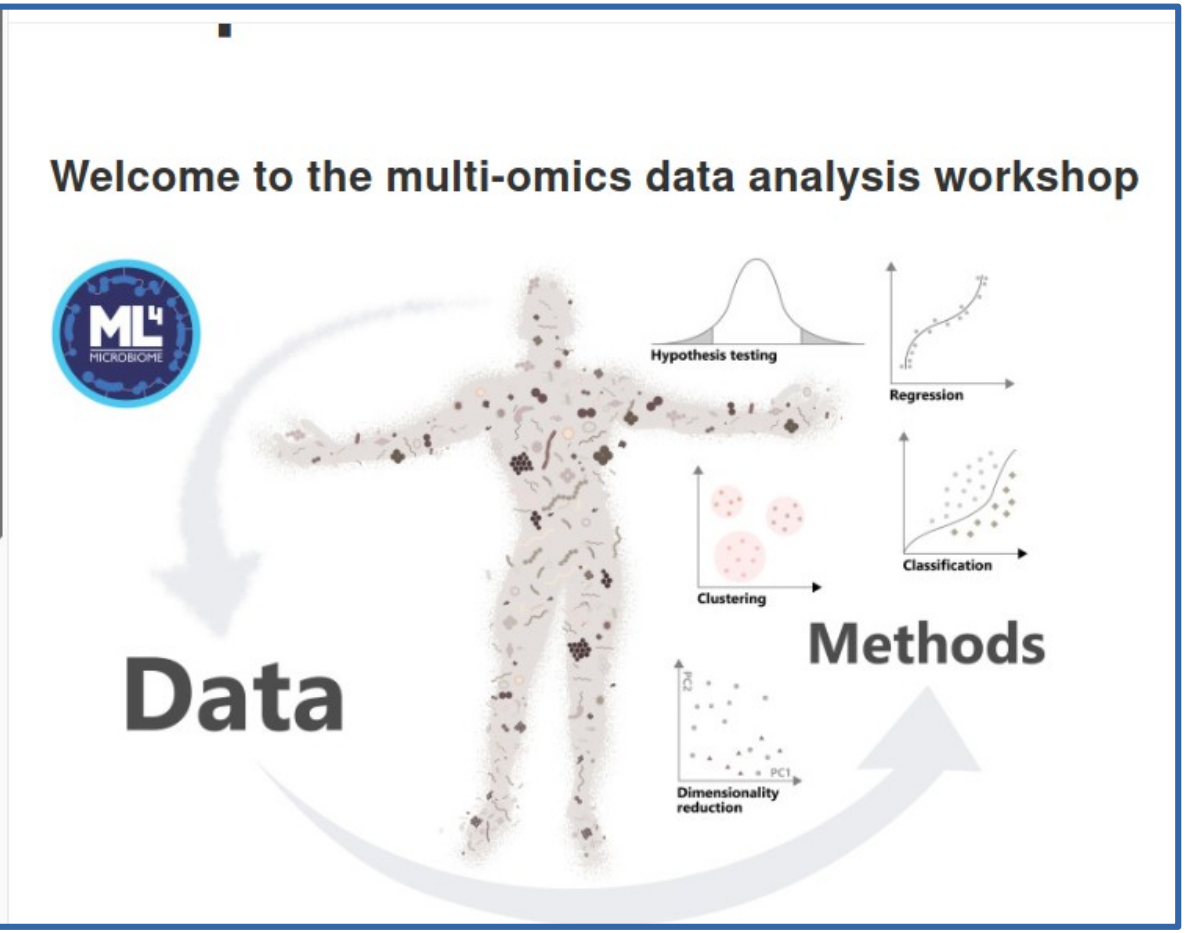


Figure source: Moreno-Indias et al. (2021) Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. Frontiers in Microbiology 12:11.

Chapter 3 Getting started

3.1 Checklist (before the workshop)

Install the following software in advance in order to avoid unnecessary delays and leaving more time for the workshop contents.

- R (version >4.1.0)
- RStudio; choose "Rstudio Desktop" to download the latest version. Optional but preferred. For further details, check the Rstudio home page.
- For Windows users: Rtools; Follow the instructions to install the toolkit. This might be required to compile some of the packages required on this course.
- Install and load the required R packages

3.2 Support and resources

For online support on installation and other matters, you can join us at:

- Users: miaverse Gitter channel
- Developers: Bioconductor Slack #microbiomeexperiment channel (ask for an invitation)

3.3 Installing and loading the required R packages

This section shows how to install and load all required packages into the R session. Only uninstalled packages are installed.