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

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

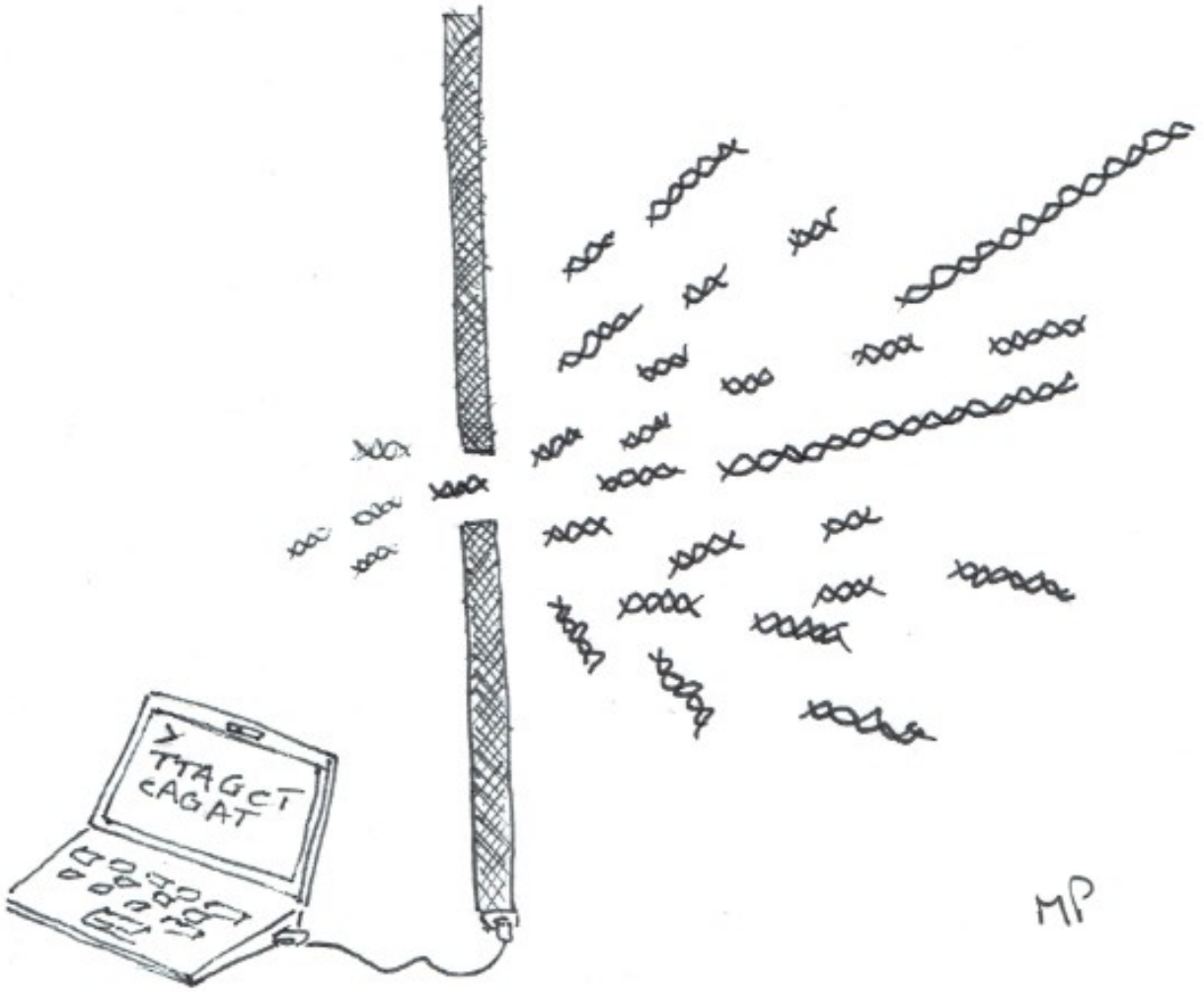
14:15-17:00 - Tuomas Borman and Chouaib Benchraka, Research assistants (UTU)

Data import and data structures

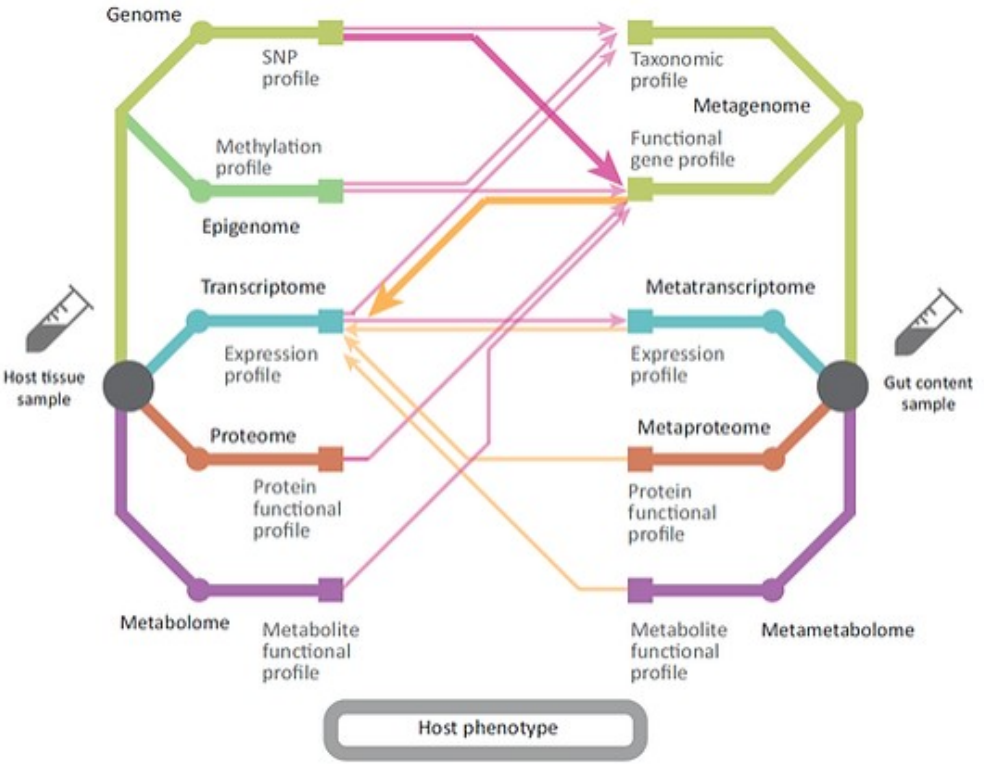
Microbiome data exploration & visualization

multi-omics

Limited observations → data integration?



By *identifying* and *integrating* biological signals in *multi-omics* data under this powerful framework, we can finally find what *causes* the rich and varied observable traits (phenotype) of a living being.



Go beyond pairwise associations towards causation

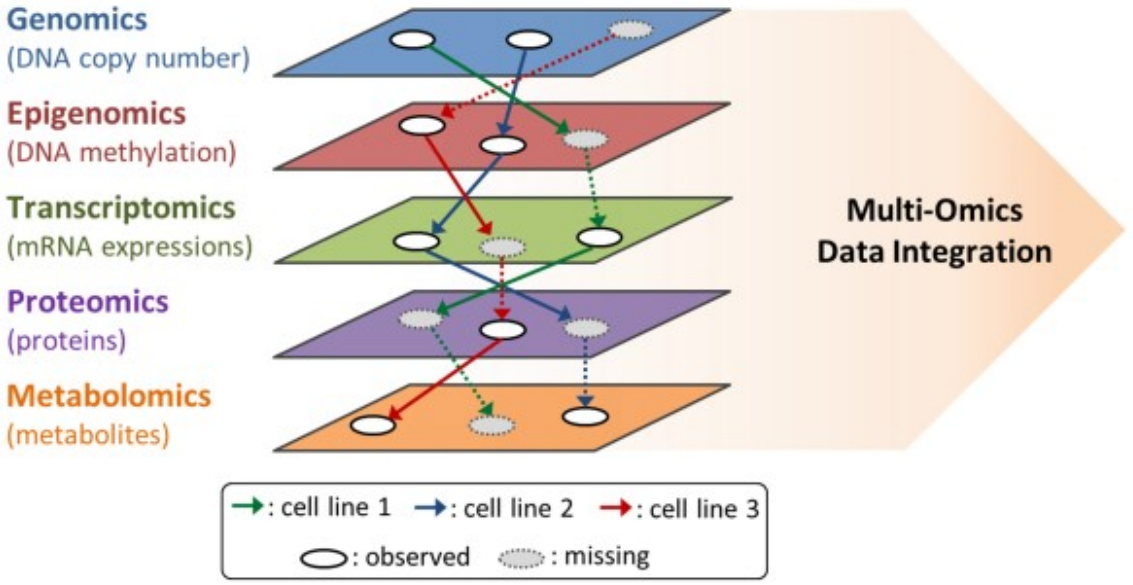
We develop methods that go beyond the current paradigm of "pairwise" associations studies by using machine learning, Bayesian statistics and causal models to determine the structure hidden in large multi-omics data sets.

Account for biological heterogeneity

We account for the true dynamic nature of the host-microbiome system by modelling both temporal and spatial changes in the microbiome and their interaction with the host environment.

Include prior knowledge

We develop new hierarchical models to incorporate external information from existing databases and research studies, such as gene or pathway information, previous association studies, and the known evolutionary consequences of genomic and metagenomic changes.



What multi-omics data integration can offer

- More comprehensive understanding of biological systems
- Improved prediction of outcomes of interests (e.g., disease traits, drug responses)

Three technical challenges:

Complex interactions

Integration of information within and across observed omics

Incomplete observations

- Observations with various omics-missing patterns
- No information loss and distortion

Cost efficiency

Value of incorporating each omics observation is unknown

https://www.vanderschaar-lab.com/self-semi-supervised-learning/

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.

Prediction / Association / Supervised learning

- Regression
- PLS-DA
- Random Forest
- SVM
- etc.

Integration of multi-omics data for prediction of phenotypic traits using random forest

Animesh Acharjee, Bjorn Kloosterman, Richard G. F. Visser & Chris Maliepaard

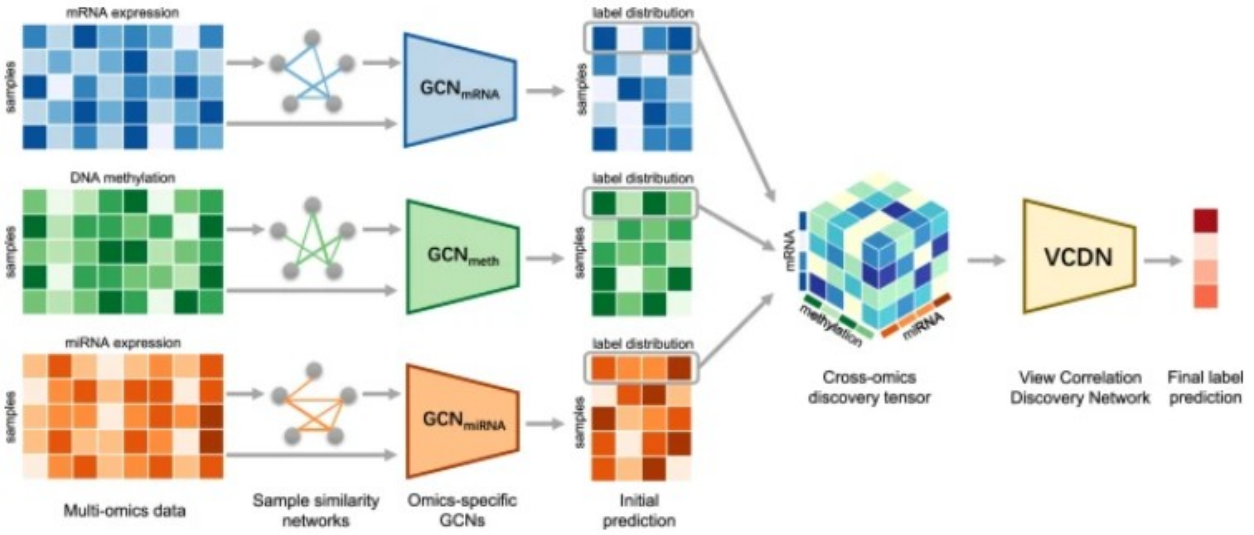
BMC Bioinformatics **17**, Article number: 180 (2016) | Cite this article

MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification

Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding & Kun Huang

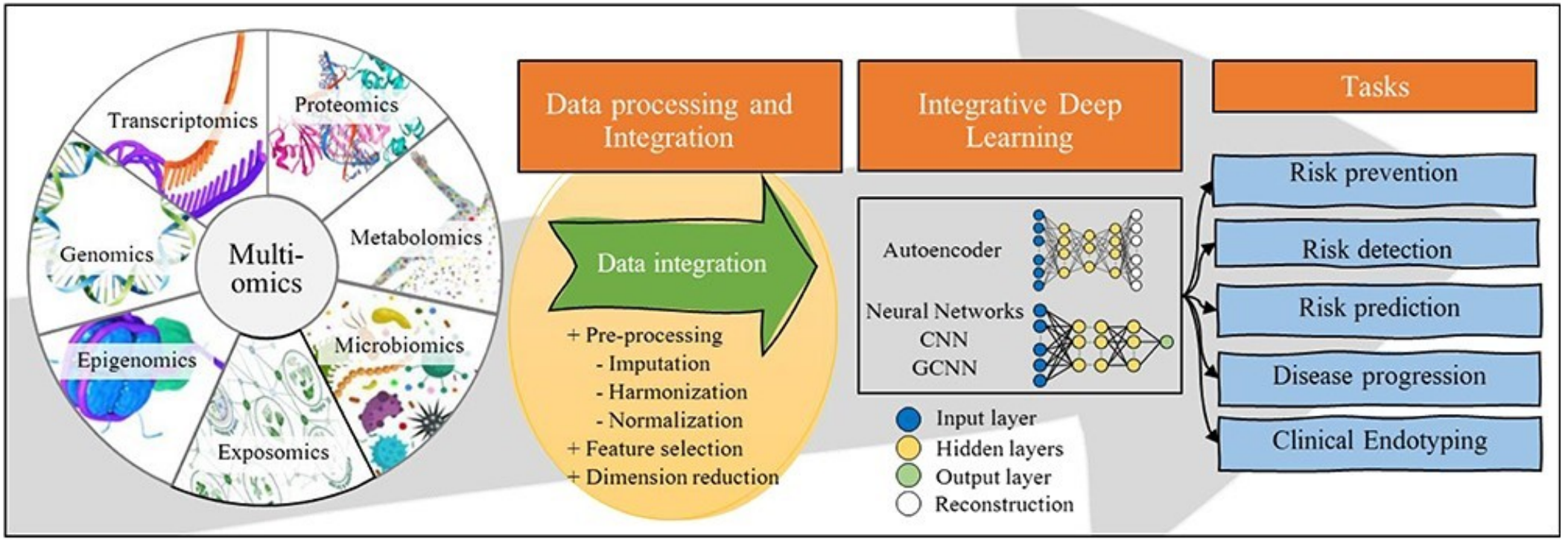
Nature Communications **12**, Article number: 3445 (2021) | Cite this article

Fig. 1: Illustration of MOGONET.



MOGONET combines GCN for multi-omics-specific learning and VCDN for multi-omics integration. For clear and concise illustration, an example of one sample is chosen to demonstrate the VCDN component for multi-omics integration. Preprocessing is first performed on each omics data type to remove noise and redundant features. Each omics-specific GCN is trained to perform class prediction using omics features and the corresponding sample similarity network generated from the omics data. The cross-omics discovery tensor is calculated from the initial predictions of omics-specific GCNs and forwarded to VCDN for final prediction. MOGONET is an end-to-end model and all networks are trained jointly.

Deep learning?



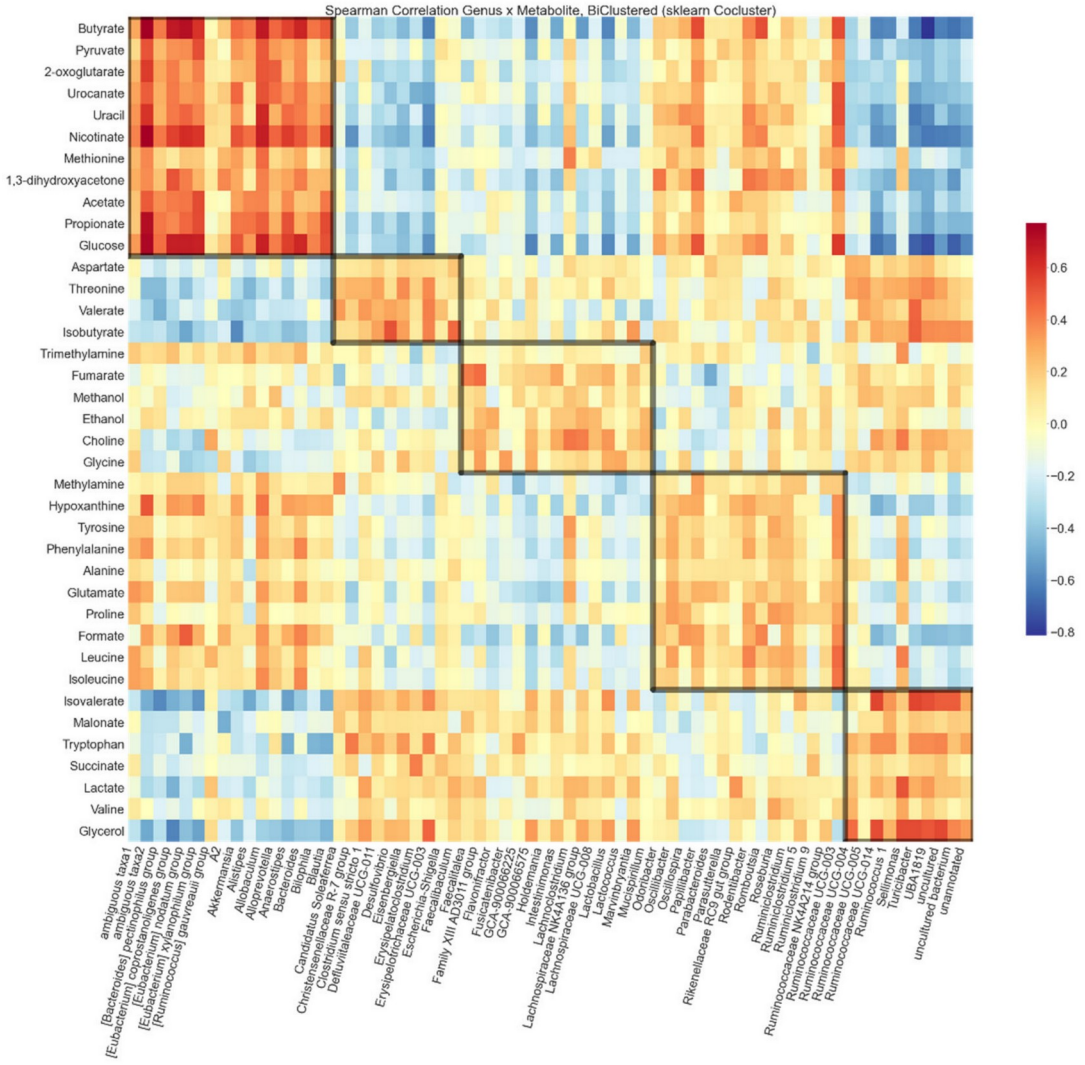
Brief Bioinform, bbab454, https://doi.org/10.1093/bib/bbab454

Bi-clustering: cross-correlating data sets (microbiota & serum metabolites)

Open Access Article

Xylo-Oligosaccharides in Prevention of Hepatic Steatosis and Adipose Tissue Inflammation: Associating Taxonomic and Metabolomic Patterns in Fecal Microbiomes with Biclustering

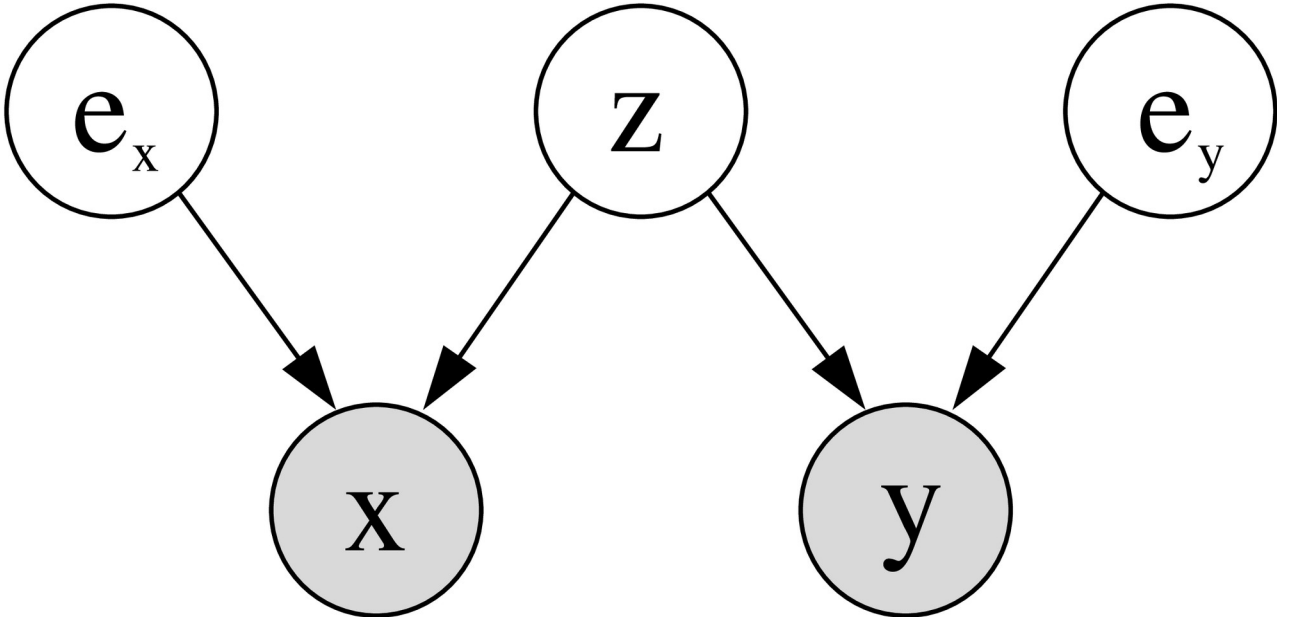
by Jukka Hintikka [1,*], Sanna Lensu [1,2], Elina Mäkinen [1], Sira Karvinen [1], Marjaana Honkanen [1], Jere Lindén [3], Tim Garrels [4], Satu Pekkala [1,5,†] and Leo Lahti [4,†]

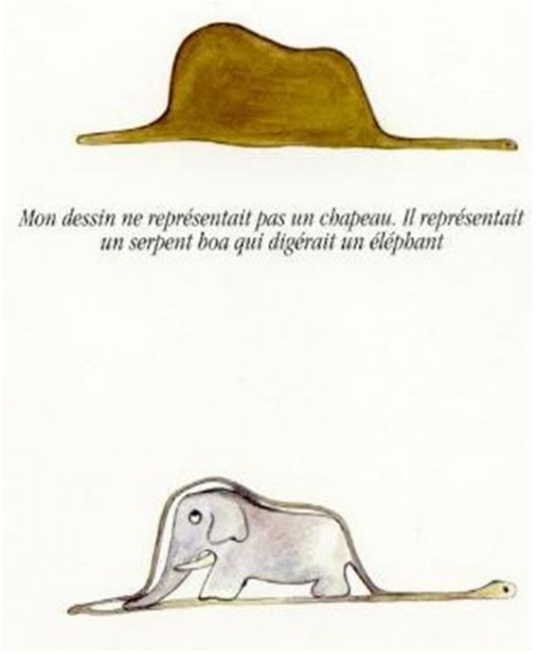


(**a**)

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

Multi-view learning





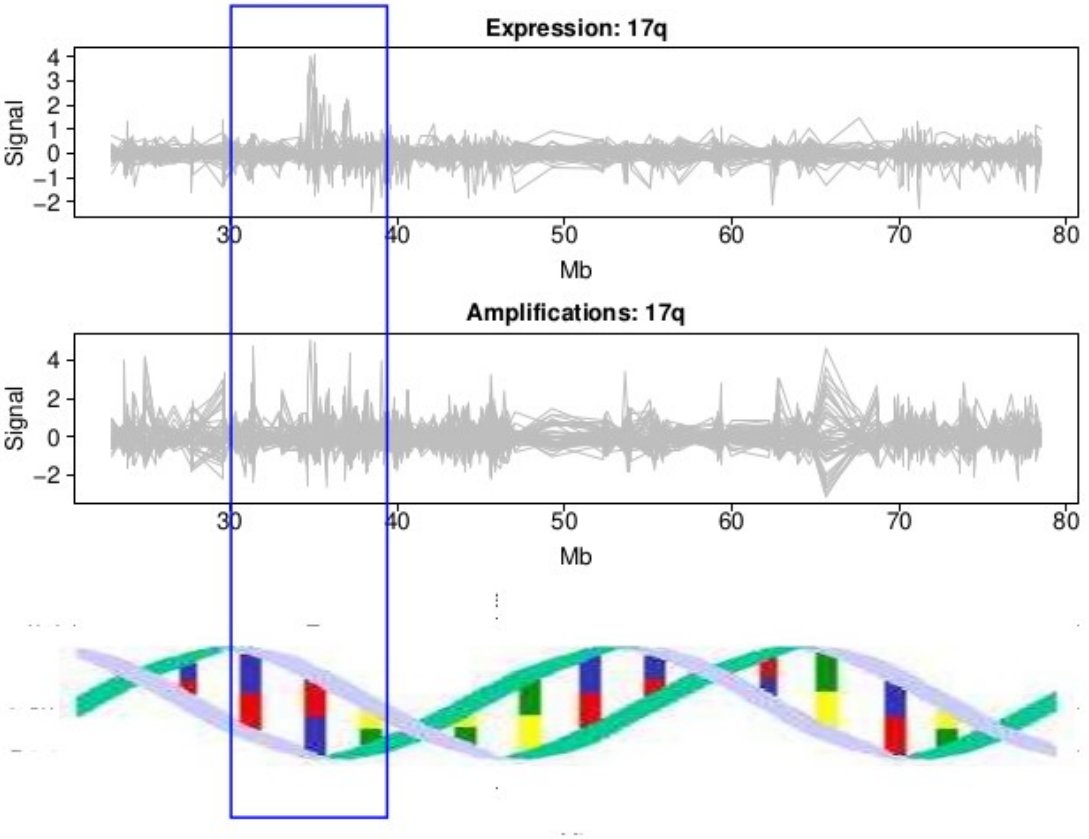
https://github.com/mblstamps/stamps2019/blob/master/STAMPS2019_overview_Pop.pdf

Chromosome arm 17q

Investigate dependencies within local chromosomal regions using sliding window

X

Y

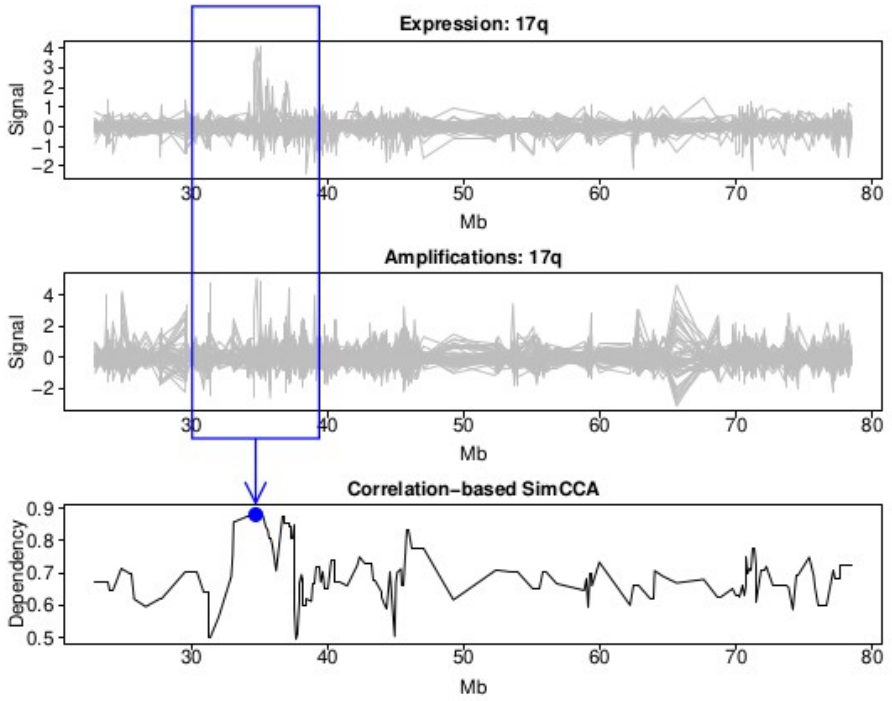


Chromosome arm 17q: results

SimCCA measures dependency between data sources within each chromosomal region

X

Y

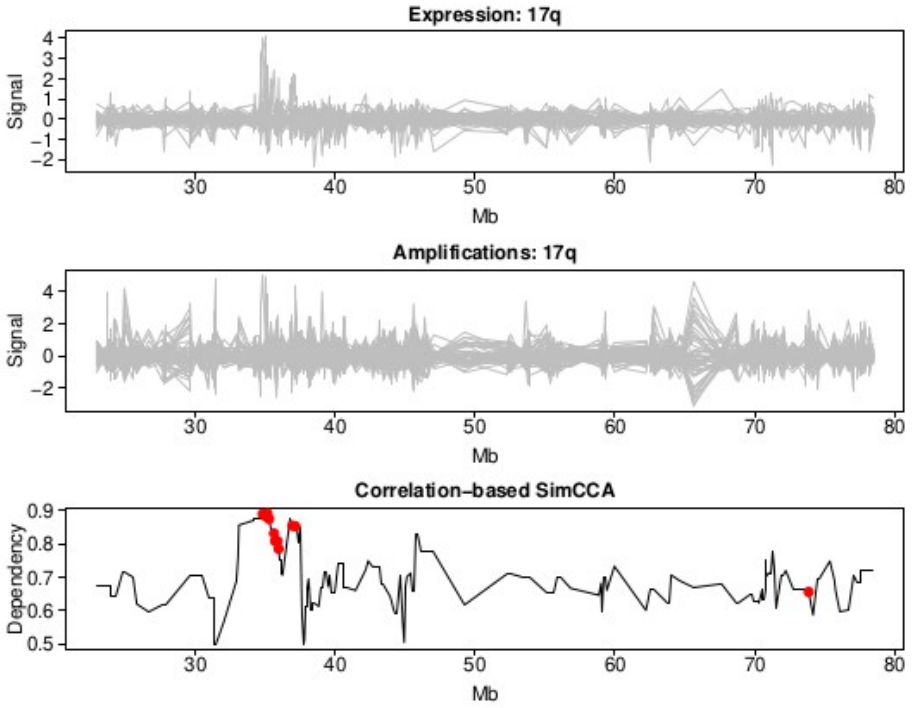


Chromosome arm 17q: results

SimCCA reveals known gastric cancer-associated chromosomal regions

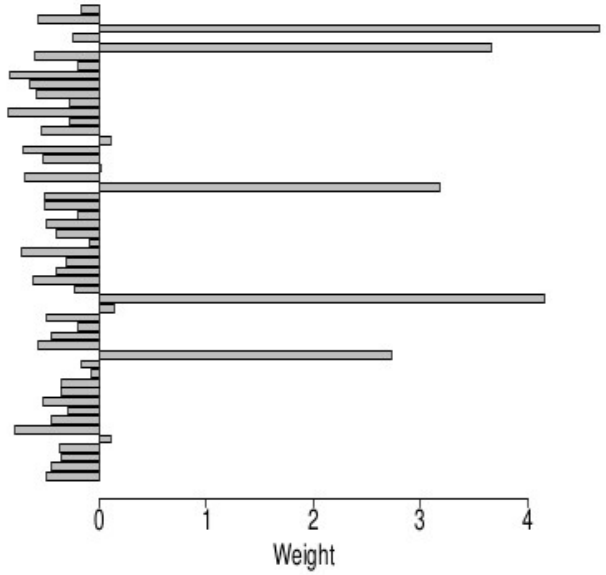
X

Y

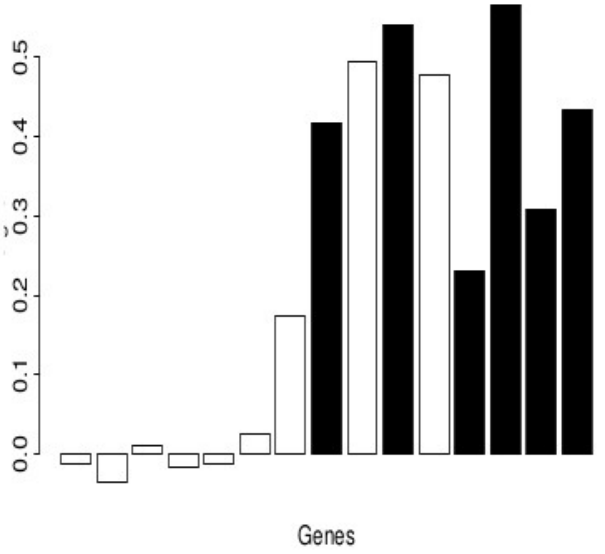


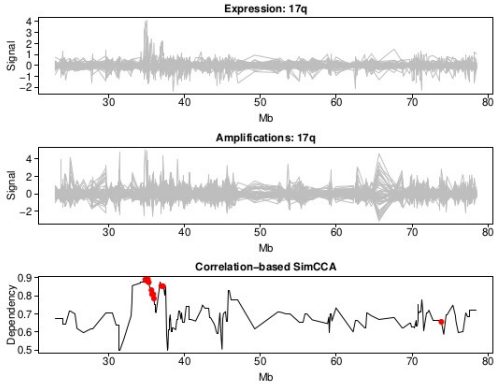
Interpreting the parameters

Z: affected patients

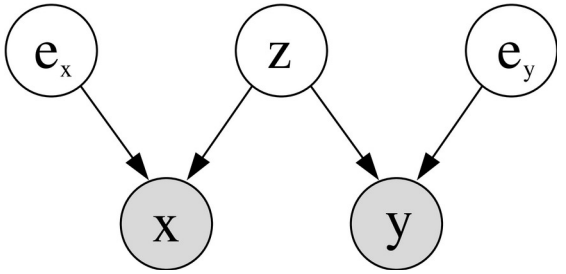


W: dependent observations





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



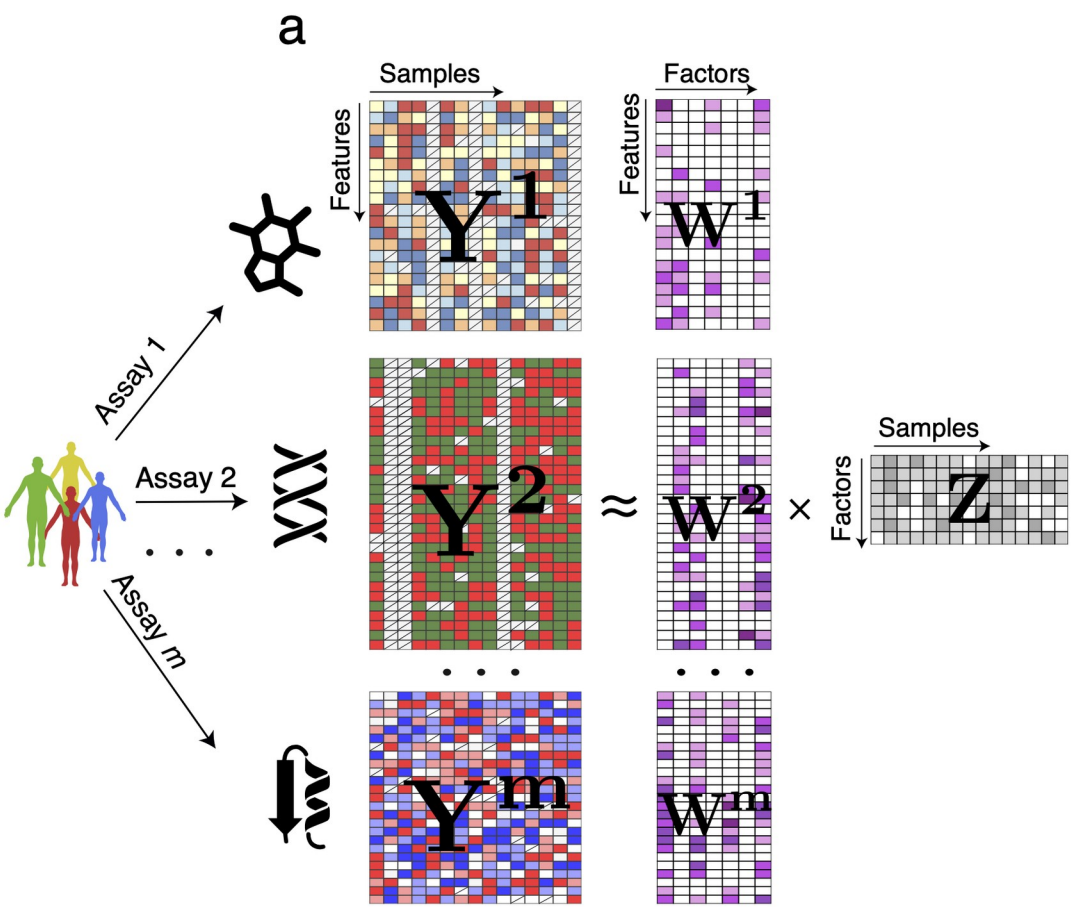
Method | 20 June 2018 | OPEN ACCESS TRANSPARENT PROCESS

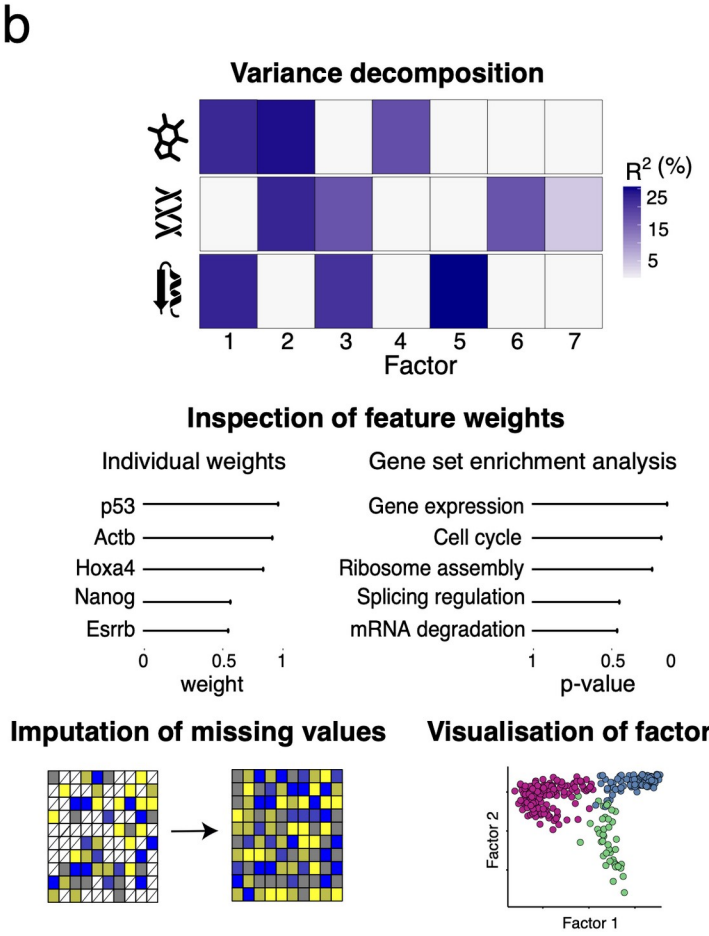
Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets

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/

Mechanisms, causality?


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

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

A survey for microbiome analysis tools in R: Github.com/microsud/Tools-Microbiome-Analysis

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

Microbiome data science

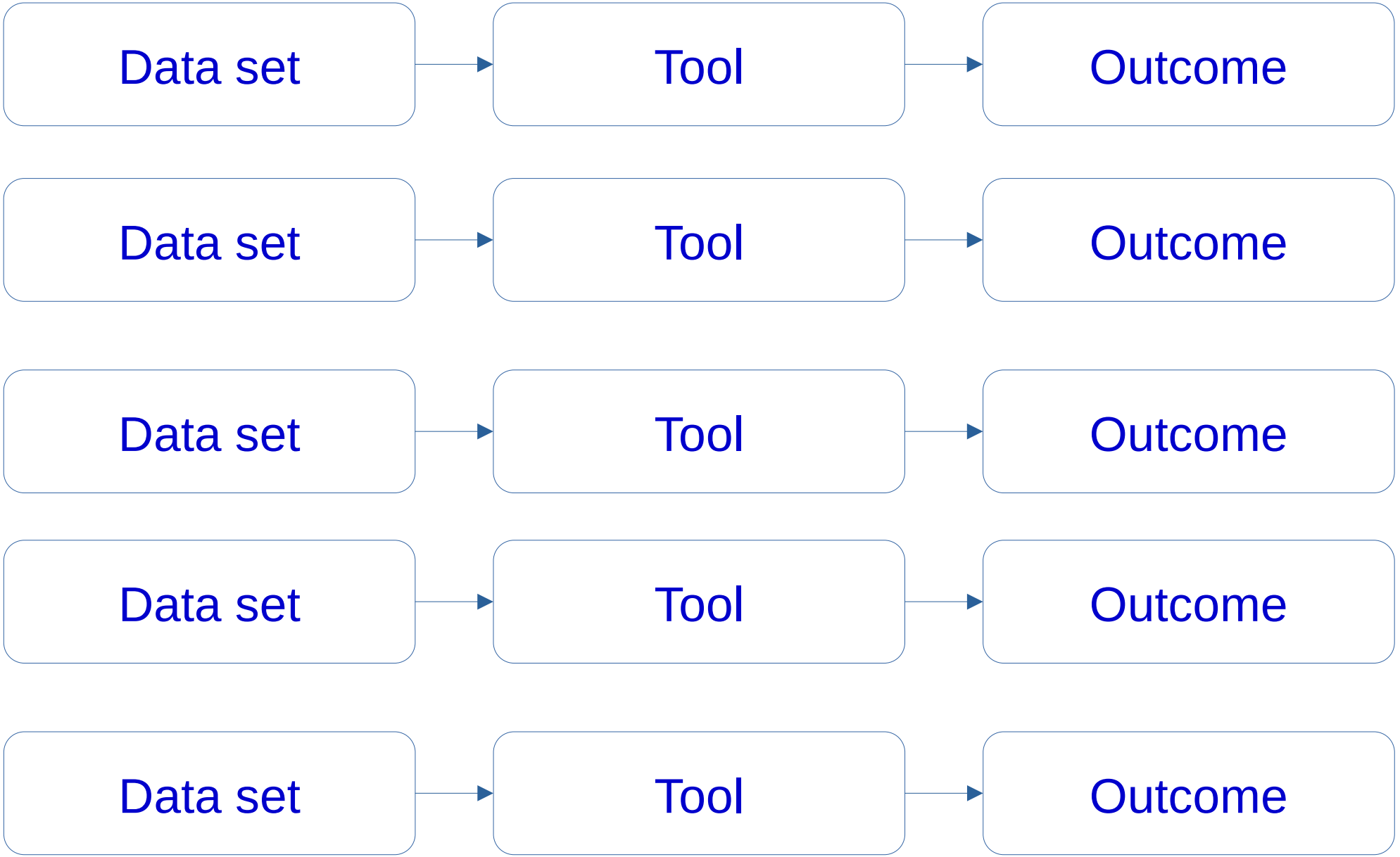
Authors | Authors and affiliations

Sudarshan A Shetty, Leo Lahti

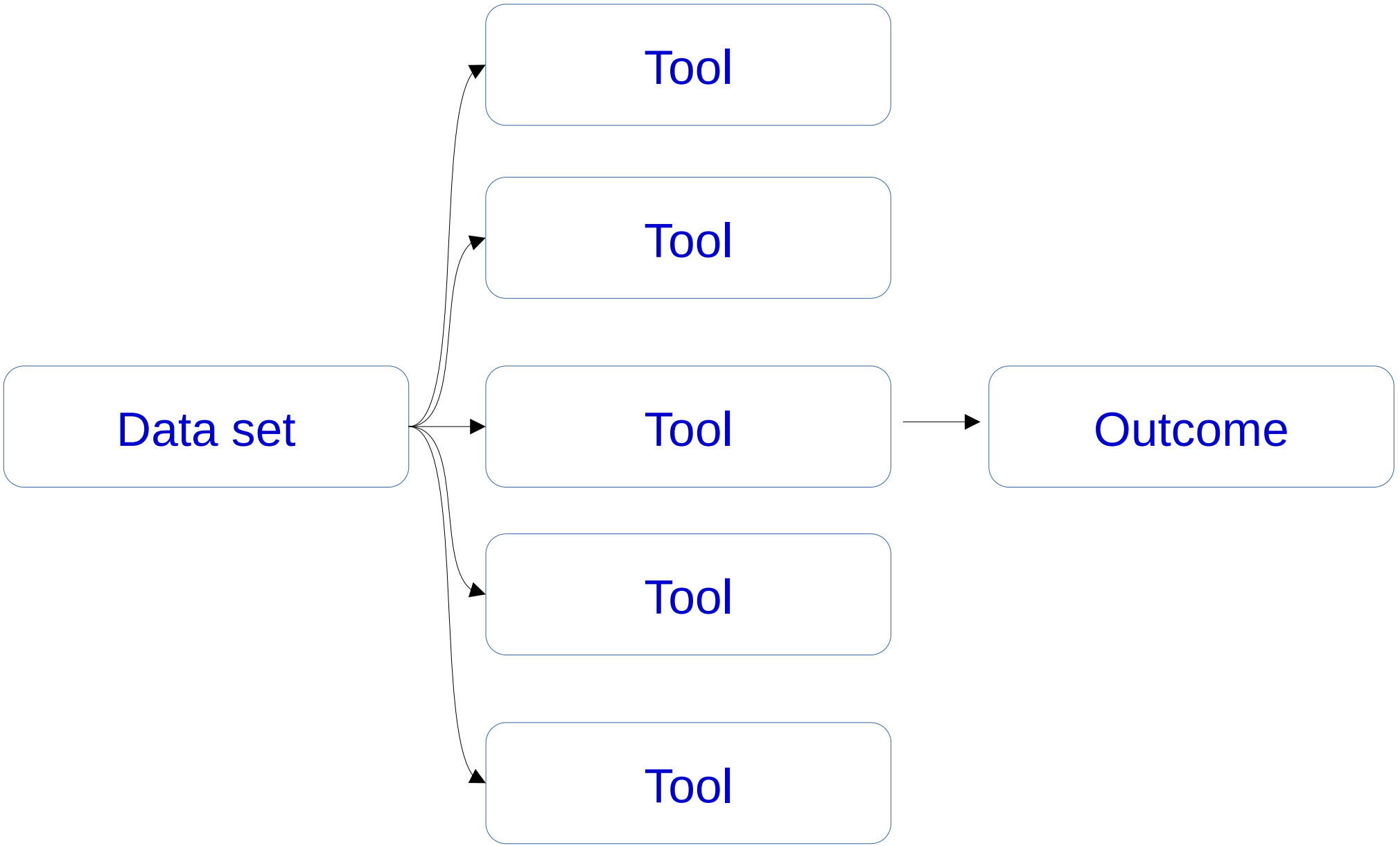
1. Ampvis2 Tools for visualising amplicon sequencing data
2. CCREPE Compositionality Corrected by PErmutation and REnormalization
3. DADA2 Divisive Amplicon Denoising Algorithm
4. DESeq2 Differential expression analysis for sequence count data
5. edgeR empirical analysis of DGE in R
6. mare Microbiota 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

→ Compatibility?

Different data, different tools?

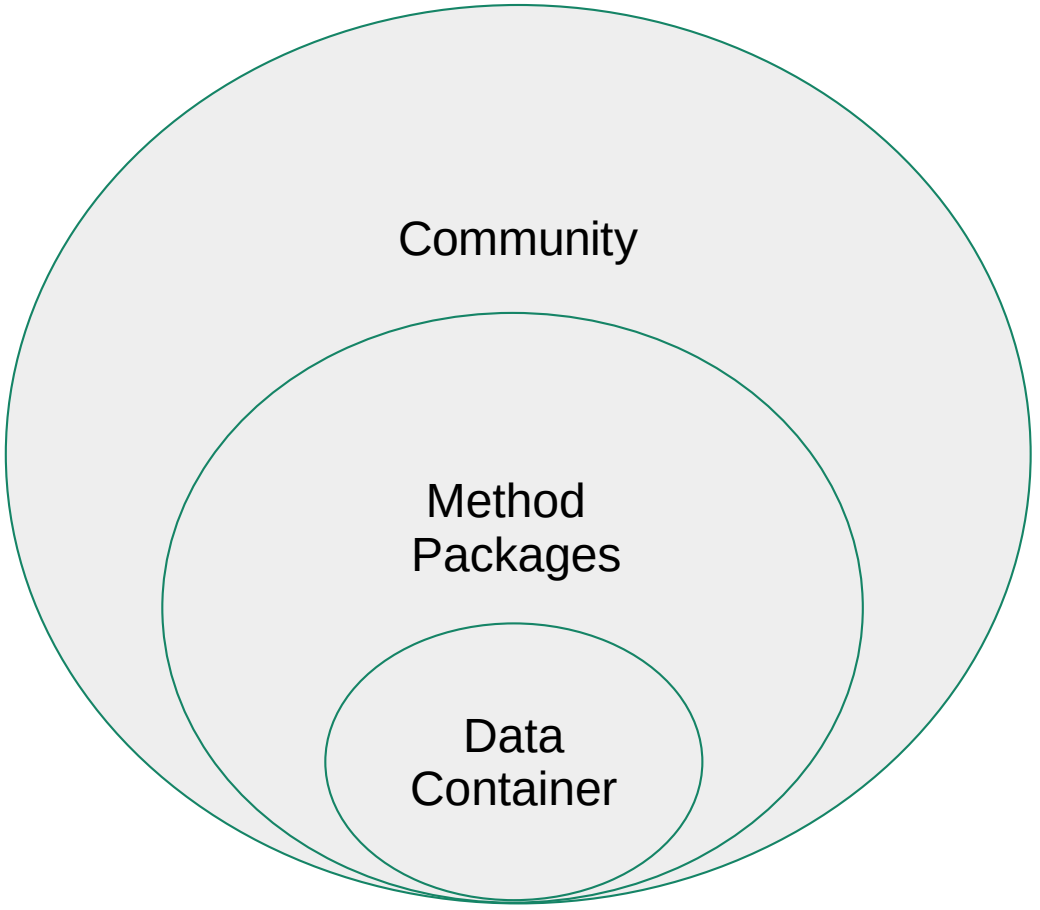


Reduce overlapping efforts, improve interoperability, ensure sustainability.



Reduce overlapping efforts, improve interoperability, ensure sustainability.

https://activeforlife.com/2020-outdoor-learning/



Optimal container for multi-omics data?

Multiple assays
seamless interlinking

Hierarchical data
supporting samples & features

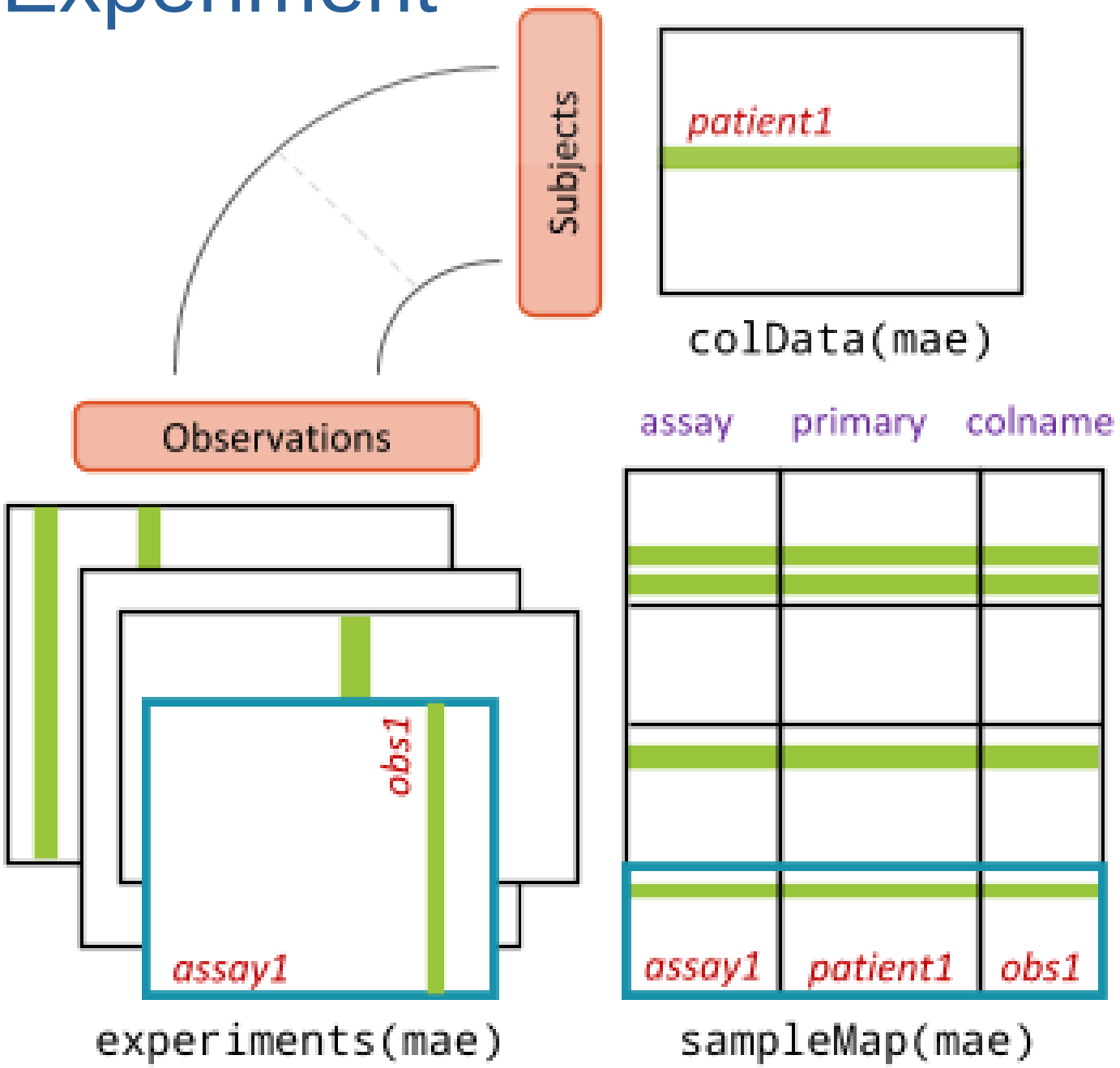
Side information
extended capabilities & data types

Optimized
for speed & memory

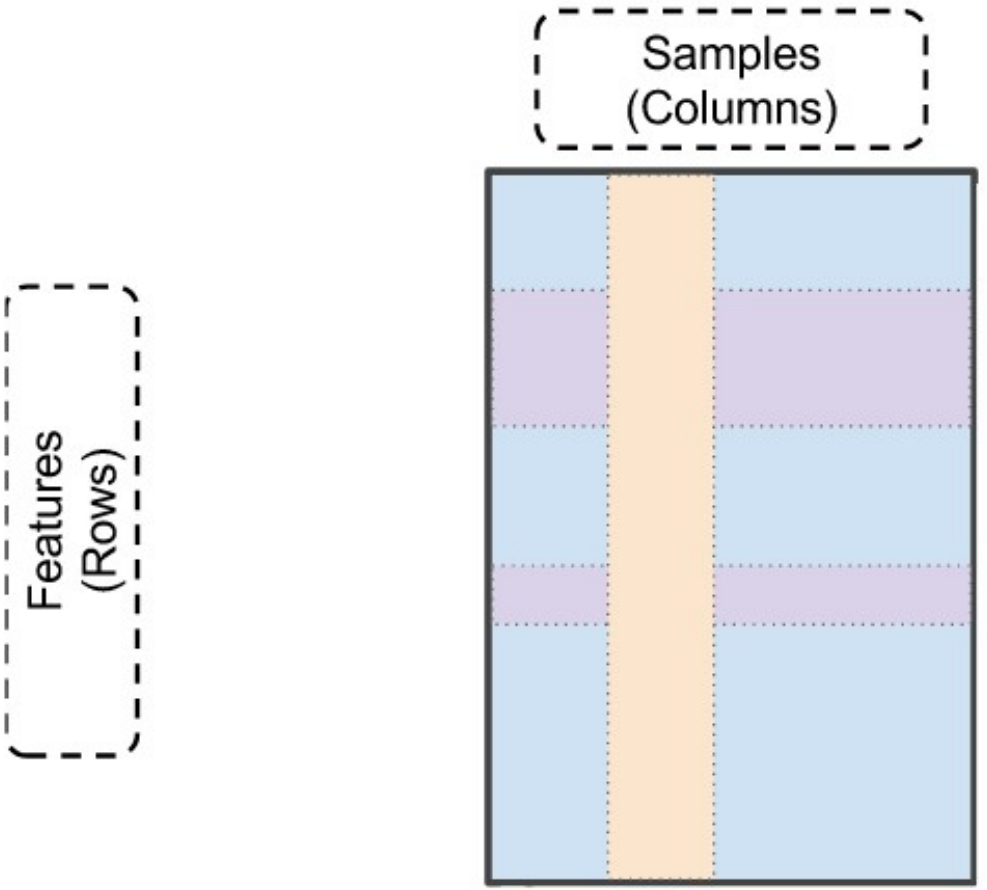
Integrated
with other applications & frameworks

Reduce overlapping efforts, improve interoperability, ensure sustainability.

MultiAssayExperiment



SummarizedExperiment



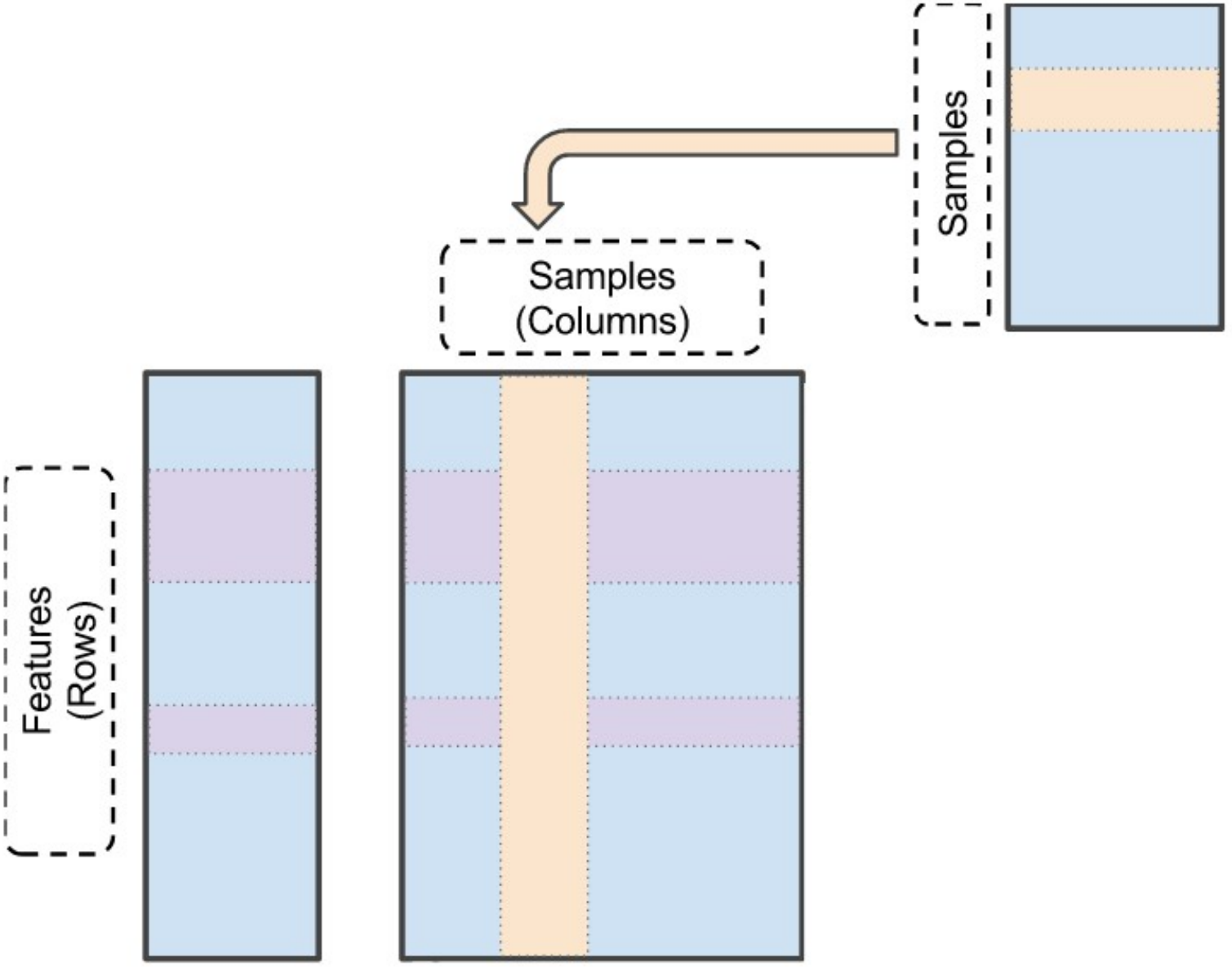
Genomics

Single Cell

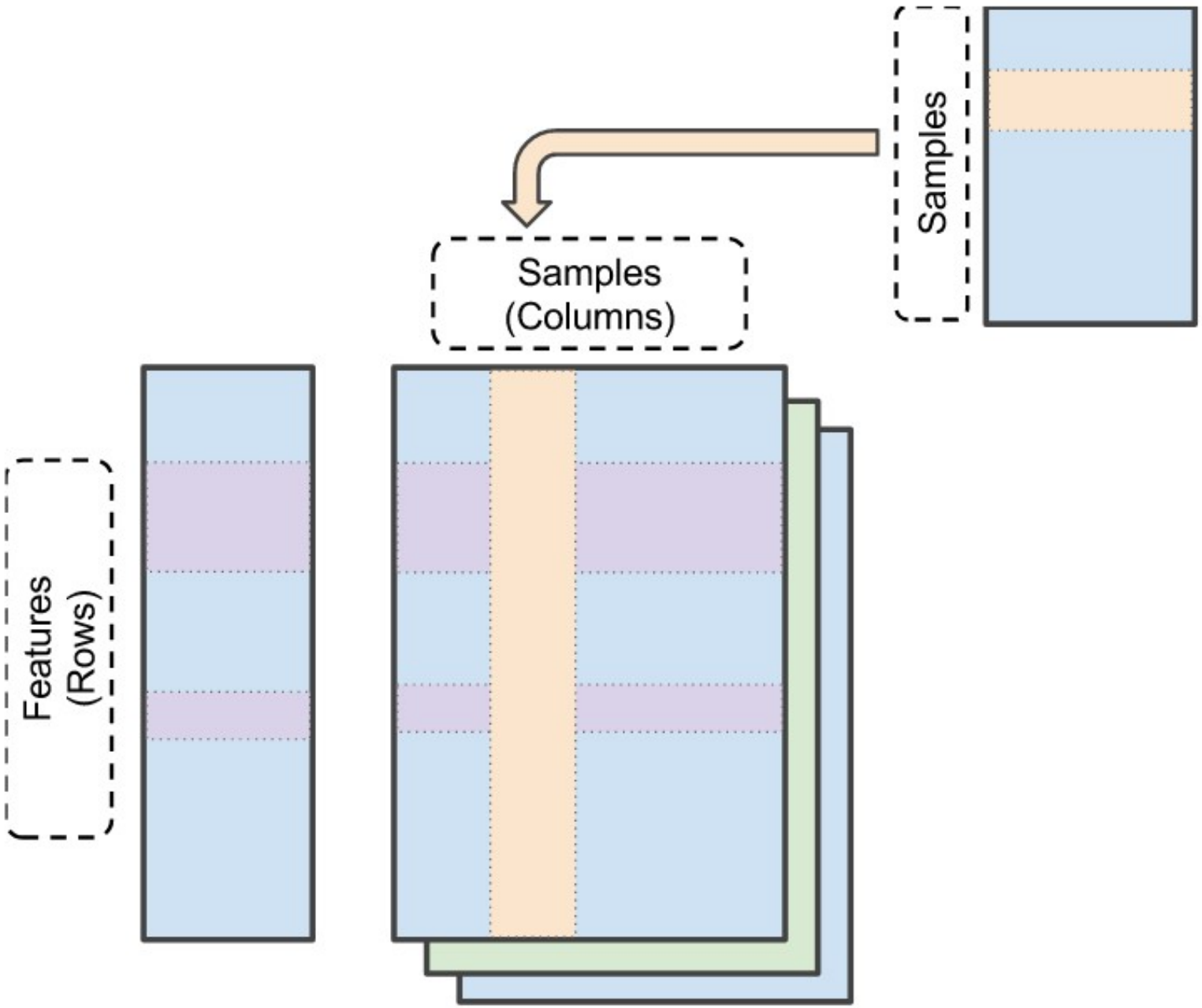
Microbiomics

...

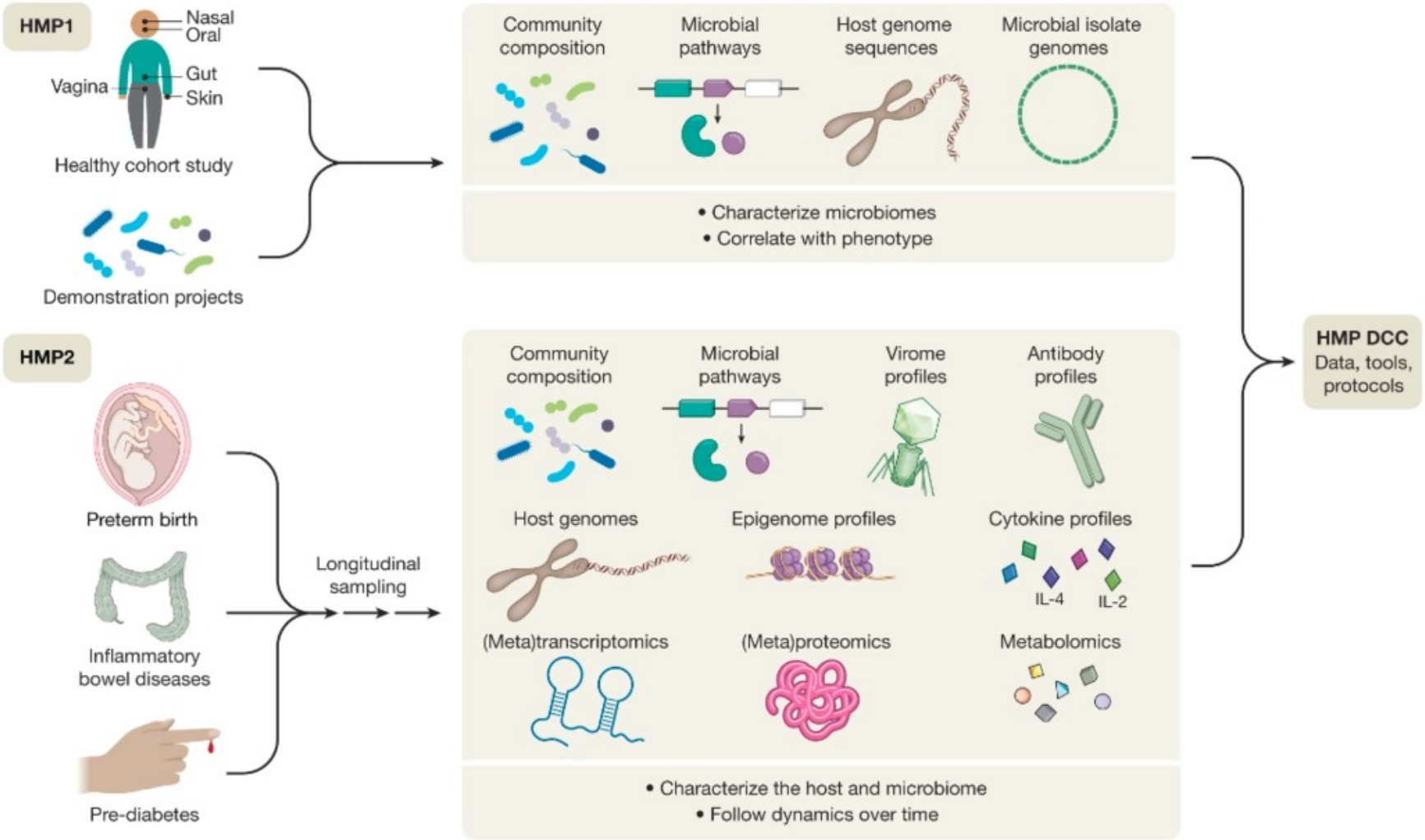
SummarizedExperiment



SummarizedExperiment

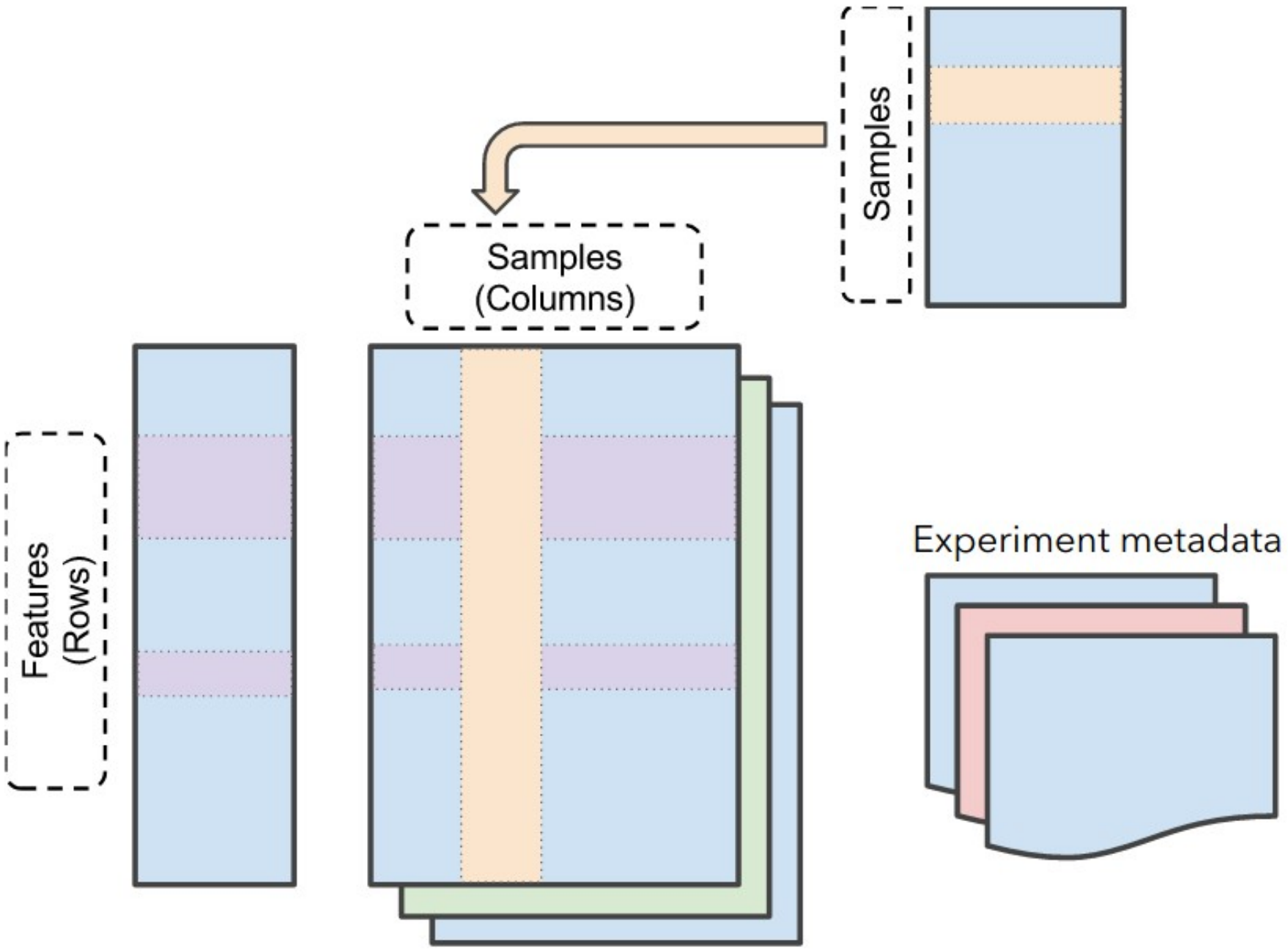


Multi-omics



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

SummarizedExperiment

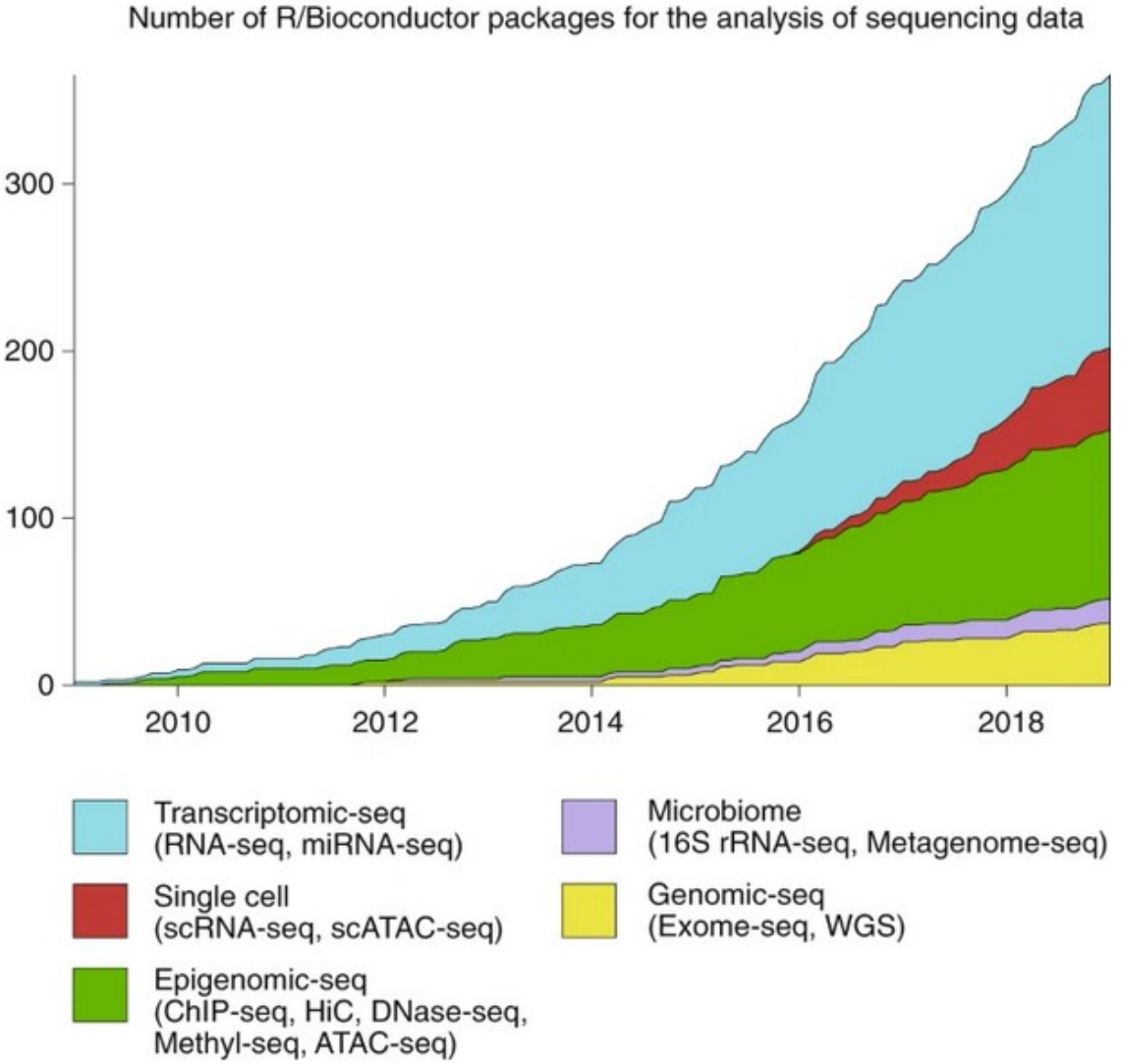


Orchestrating single-cell analysis with Bioconductor

Robert A. Amezquita, Aaron T. L. Lun, Etienne Becht, Vince J. Carey, Lindsay N. Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, Levi Waldron, Hervé Pagès, Mike L. Smith, Wolfgang Huber, Martin Morgan, Raphael Gottardo ✉ & Stephanie C. Hicks ✉

Nature Methods **17**, 137–145 (2020) | Cite this article

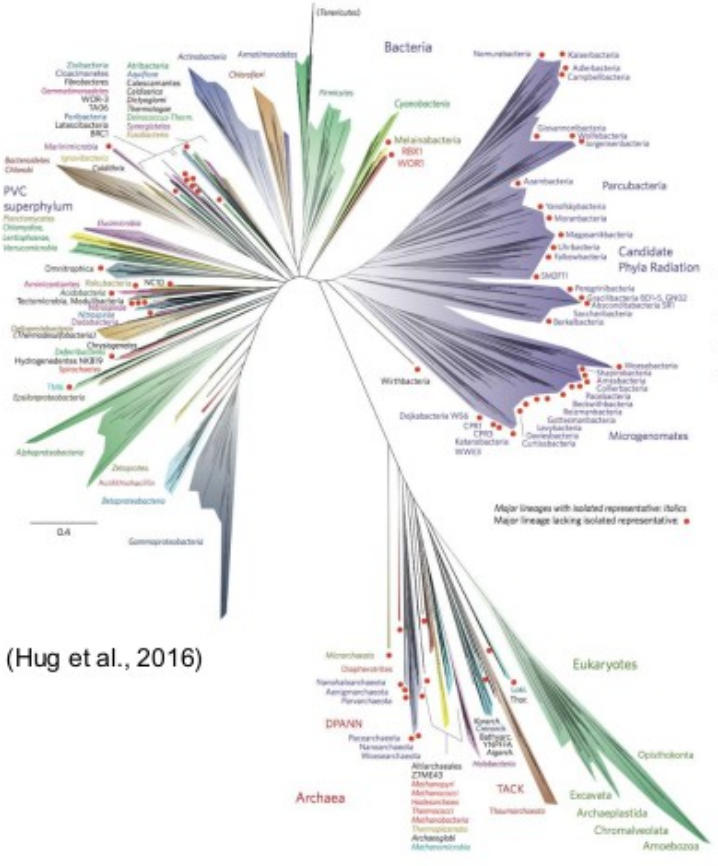
Fig. 1: Number of Bioconductor packages for the analysis of high-throughput sequencing data over ten years.



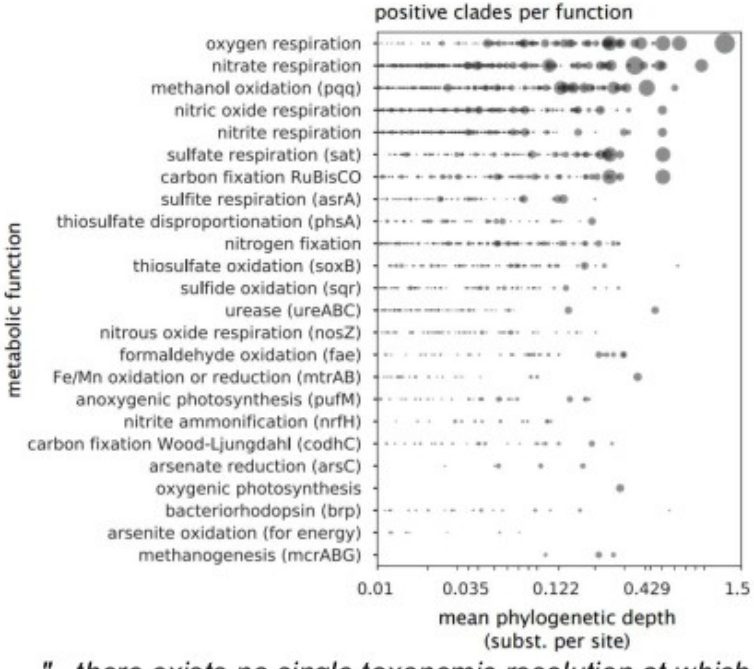
Bioconductor software packages associated with the analysis of sequencing data were tracked by date of submission over the course of ten years. Software packages were uniquely defined by their primary sequencing technology association, with examples of specific terms used for annotation in parentheses.

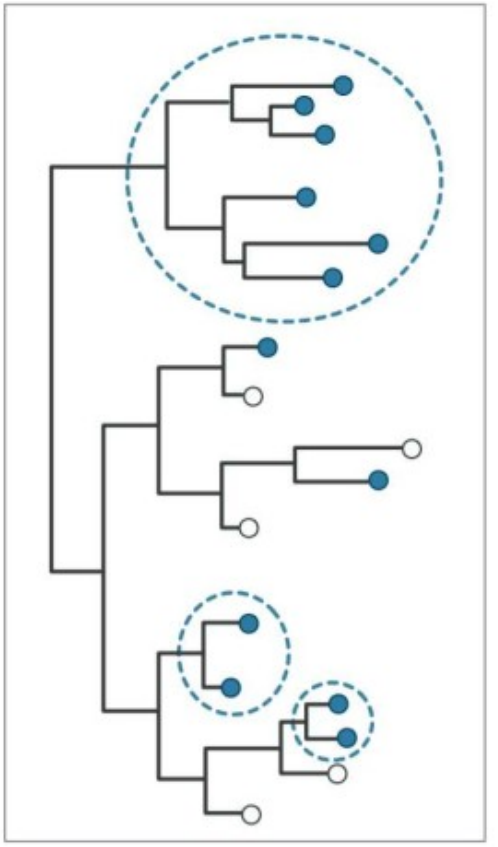
Source data

The use of phylogenetic information in metagenomics



(Hug et al., 2016)



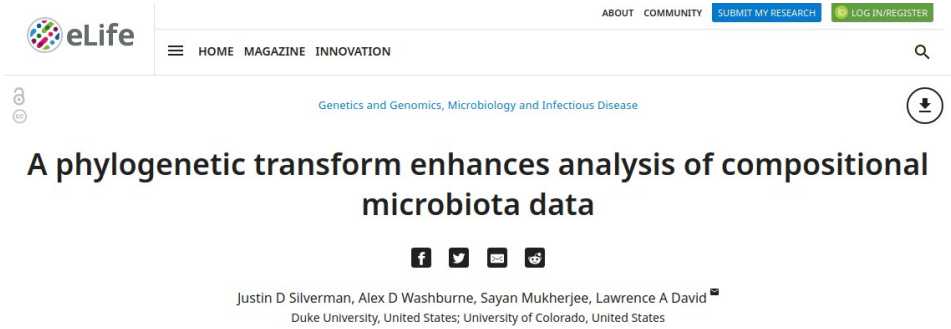


"...there exists no single taxonomic resolution at which taxonomic variation unambiguously reflects functional variation, and at which environmental selection of certain functions ... unambiguously translates to a selection of specific taxa."

(Louca et al., 2018)

TURUN YLIOPISTO

Details for FLI and PhILR transform

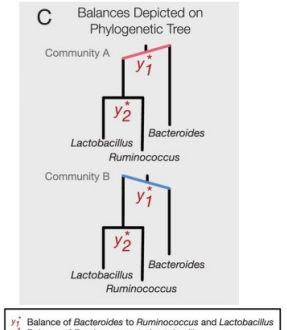


A phylogenetic transform enhances analysis of compositional microbiota data

Justin D Silverman, Alex D Washburne, Sayan Mukherjee, Lawrence A David
Duke University, United States; University of Colorado, United States

$$\mathrm{FLI} = \left(\begin{array}{l} e^{0.953 \times \log_e(\mathrm{TG}) + 0.139 \times \mathrm{BMI} + 0.718 \times \log_e(\mathrm{GGT})} \\ + 0.053 \times \text{waist circumference} - 15.745 \end{array} \right) \Big/ \left(\begin{array}{l} 1 + e^{0.953 \times \log_e(\mathrm{TG}) + 0.139 \times \mathrm{BMI} + 0.718 \times \log_e(\mathrm{GGT})} \\ + 0.053 \times \text{waist circumference} - 15.745 \end{array} \right) \times 100$$

(Bedogni et al., 2006)



(Silverman et al., 2017)

TURUN YLIOPISTO

TreeSummarizedExperiment data container

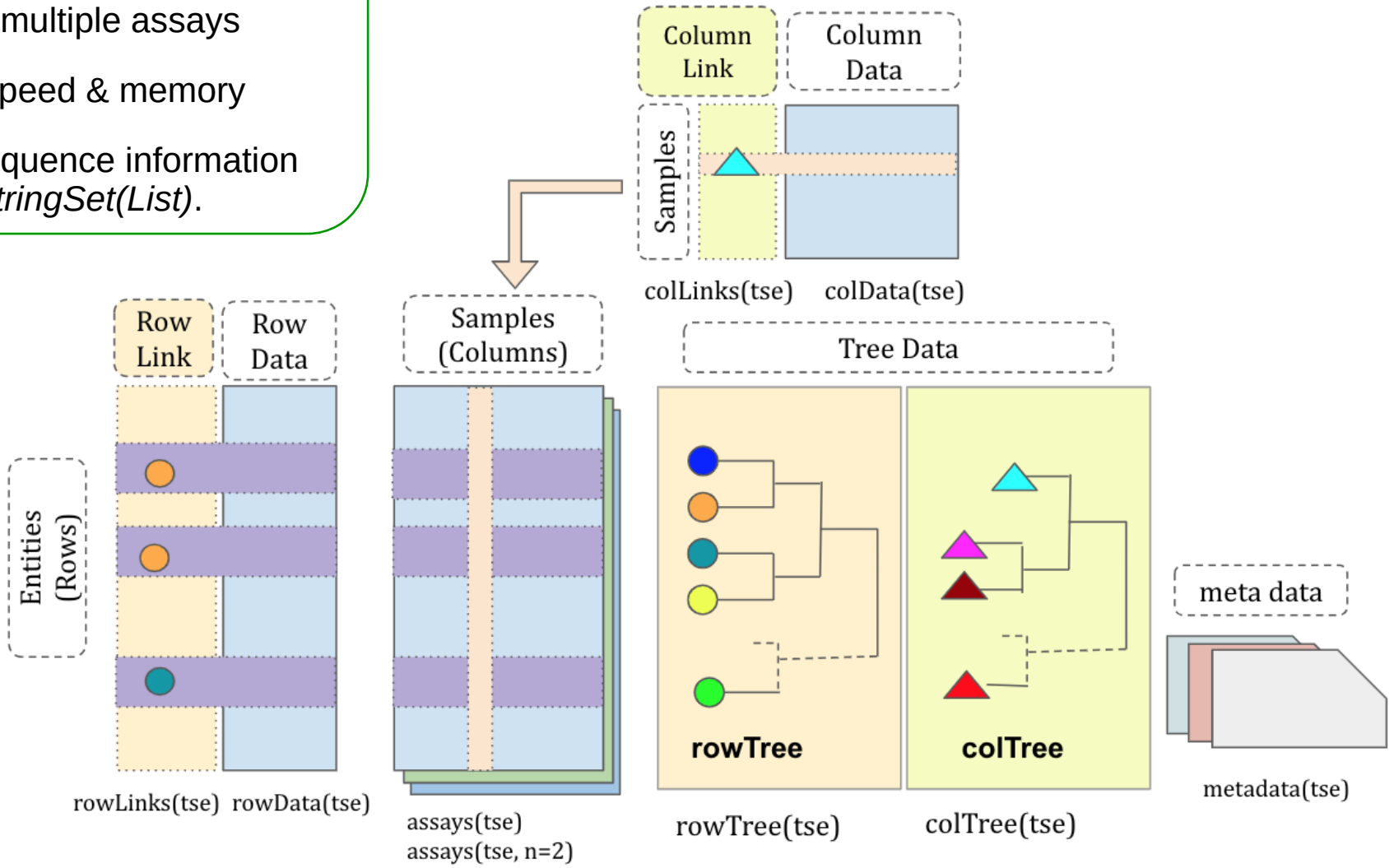
by Ruizhu @fiona Huang; initially proposed for microbiome research by Hector Bravo & Domenick Braccia

Tested tools for hierarchical data

Inherit support for sparse matrices & multiple assays

Improved speed & memory

Detailed sequence information with *DNAStringSet(List)*.

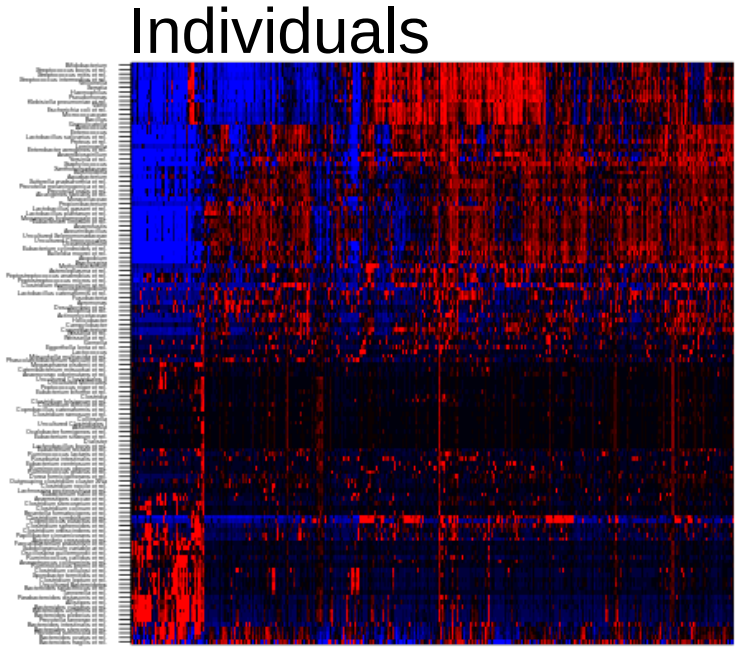


Seamless conversion from *phyloseq* & other raw data types

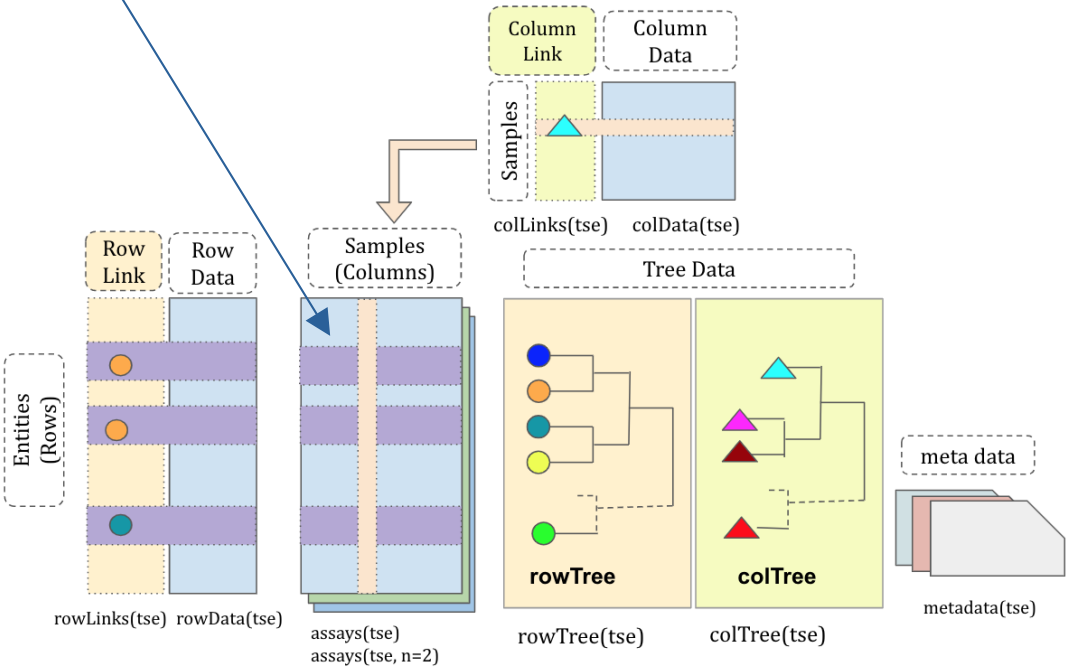
Omics data container

TreeSummarizedExperiment

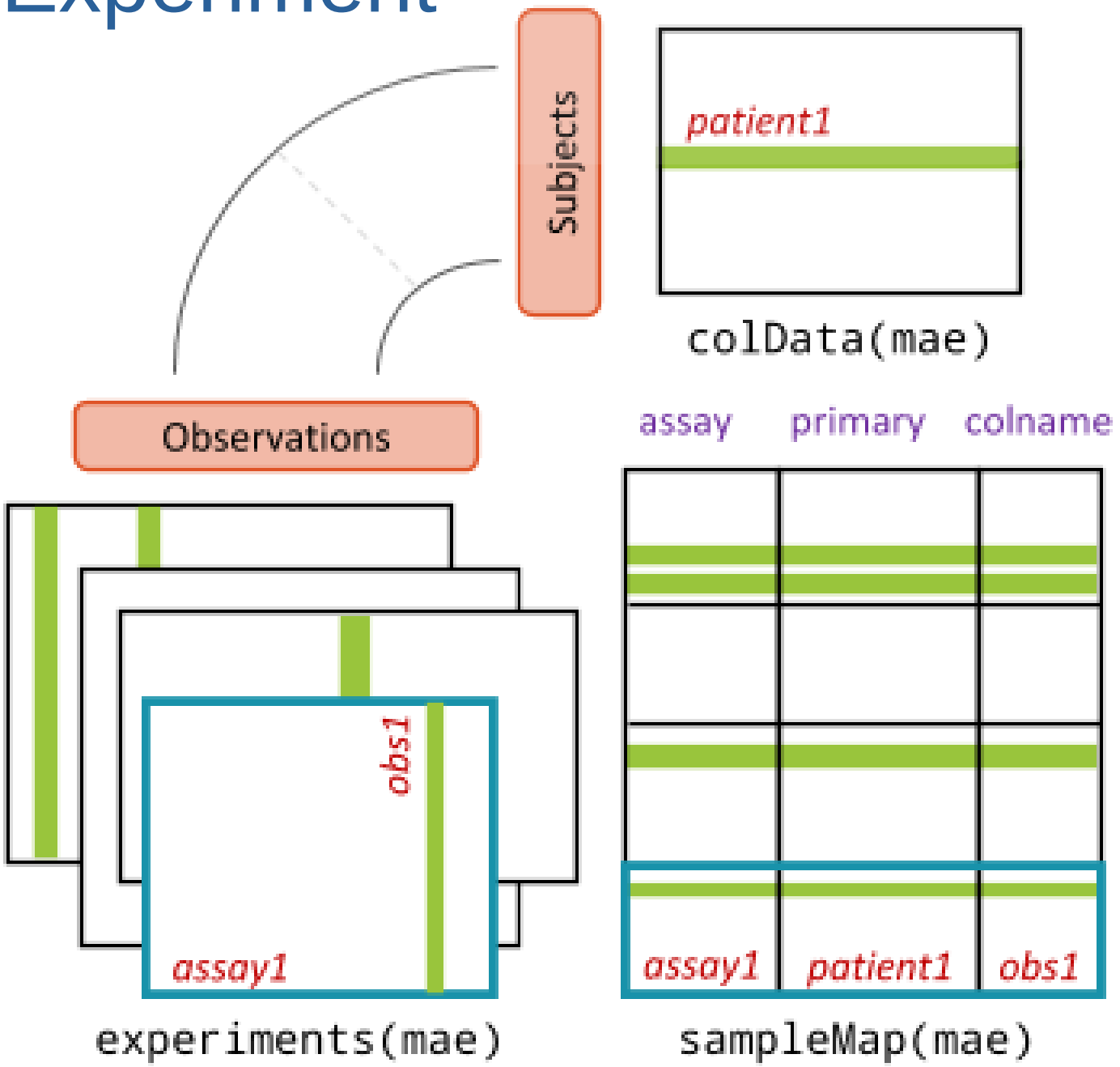
Genomics
Epigenomics
Microbiomics
Lipidomics
Proteomics
Glycomics
Foodomics
Transcriptomics
Metabolomics
Culturomics

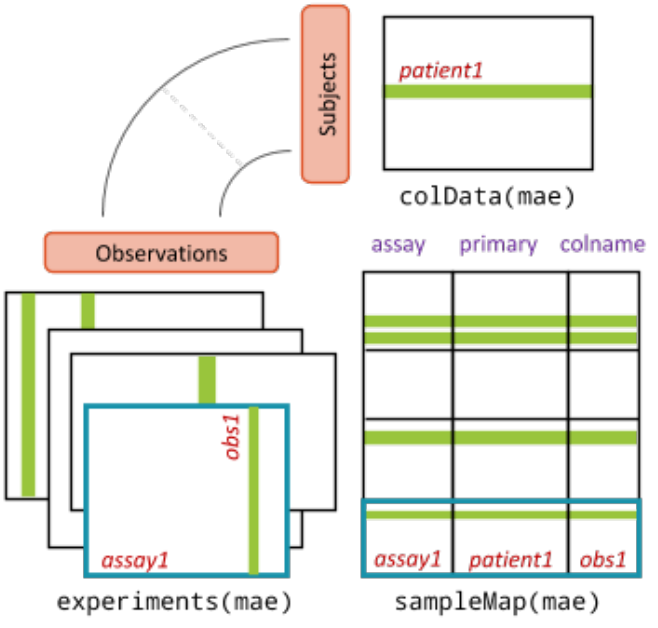


Gut microbiota: 1000 western adult
(Lahti *et al.* Nature Comm. 2014)



MultiAssayExperiment

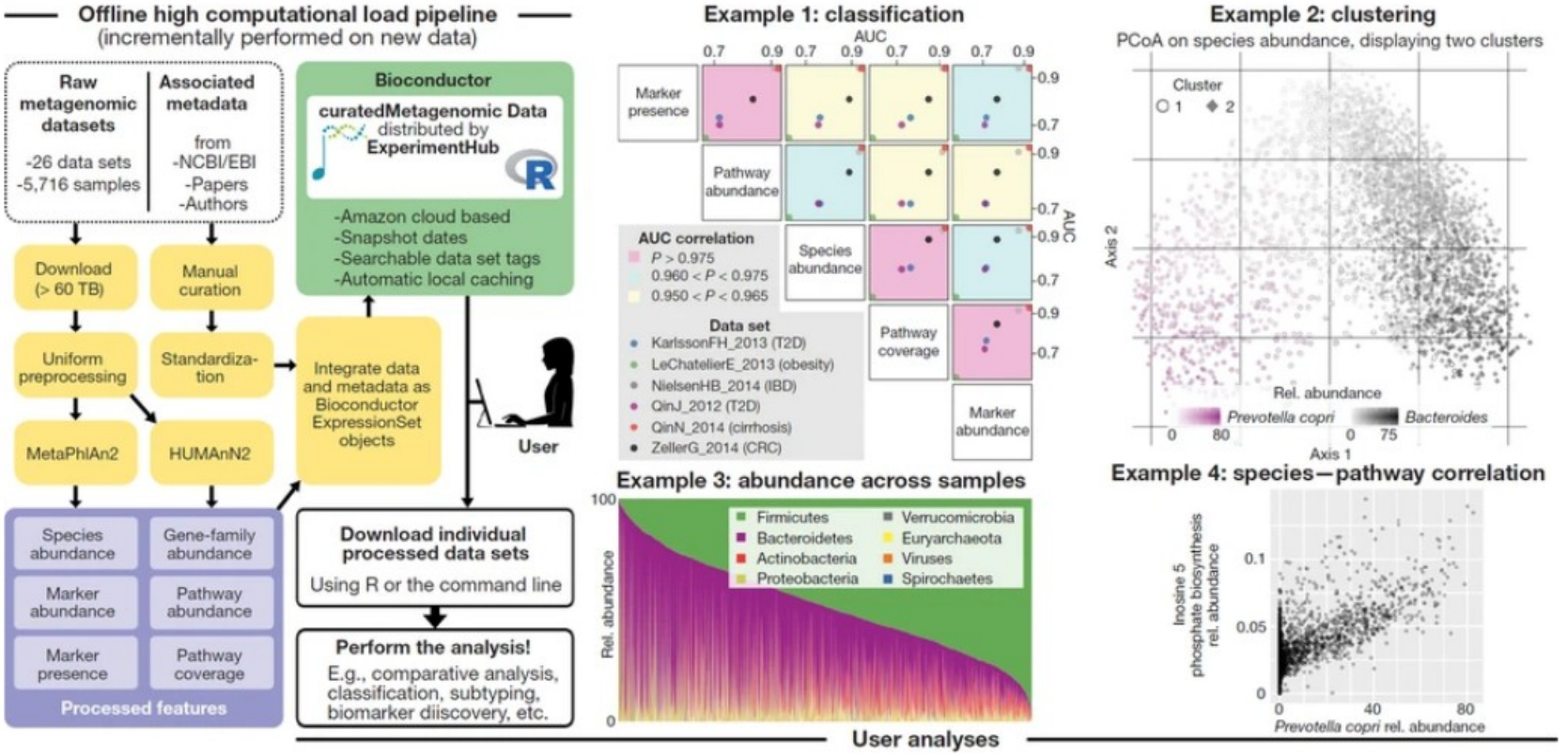




Accessible, curated metagenomic data through ExperimentHub

Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B Dowd, Curtis Huttenhower, Martin Morgan, Nicola Segata ✉ & Levi Waldron ✉

Nature Methods **14**, 1023–1024 (2017) | Cite this article



curatedMetagenomicData

platforms all | rank 30 / 408 | support 1 / 1 | build ok | updated < 1 month | dependencies 155

DOI: 10.18129/B9.bioc.curatedMetagenomicData

Curated Metagenomic Data of the Human Microbiome

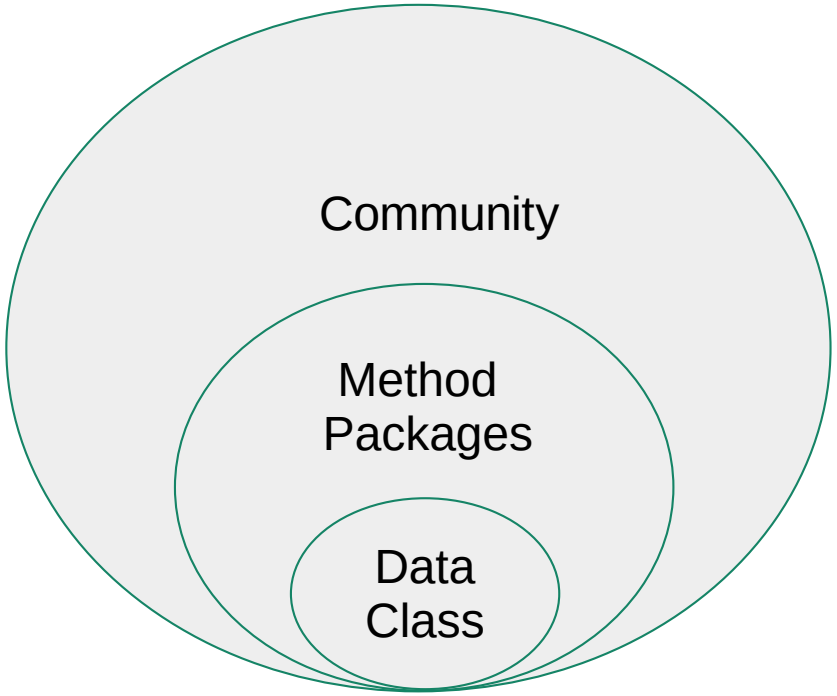
microbiomeDataSets

platforms all | rank 99 / 408 | support 0 / 0 | build ok | updated before release | dependencies 113

DOI: 10.18129/B9.bioc.microbiomeDataSets

Experiment Hub based microbiome datasets

Reduce overlapping efforts, improve interoperability, ensure sustainability.



Data packages

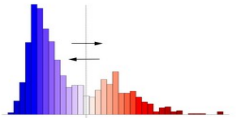
ExperimentHub

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

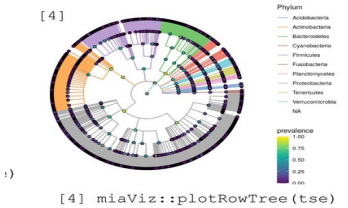
DOI: 10.18129/B9.bioc.ExperimentHub

mia – microbiome analysis

getDiversity(x)
calculateDMM(x)



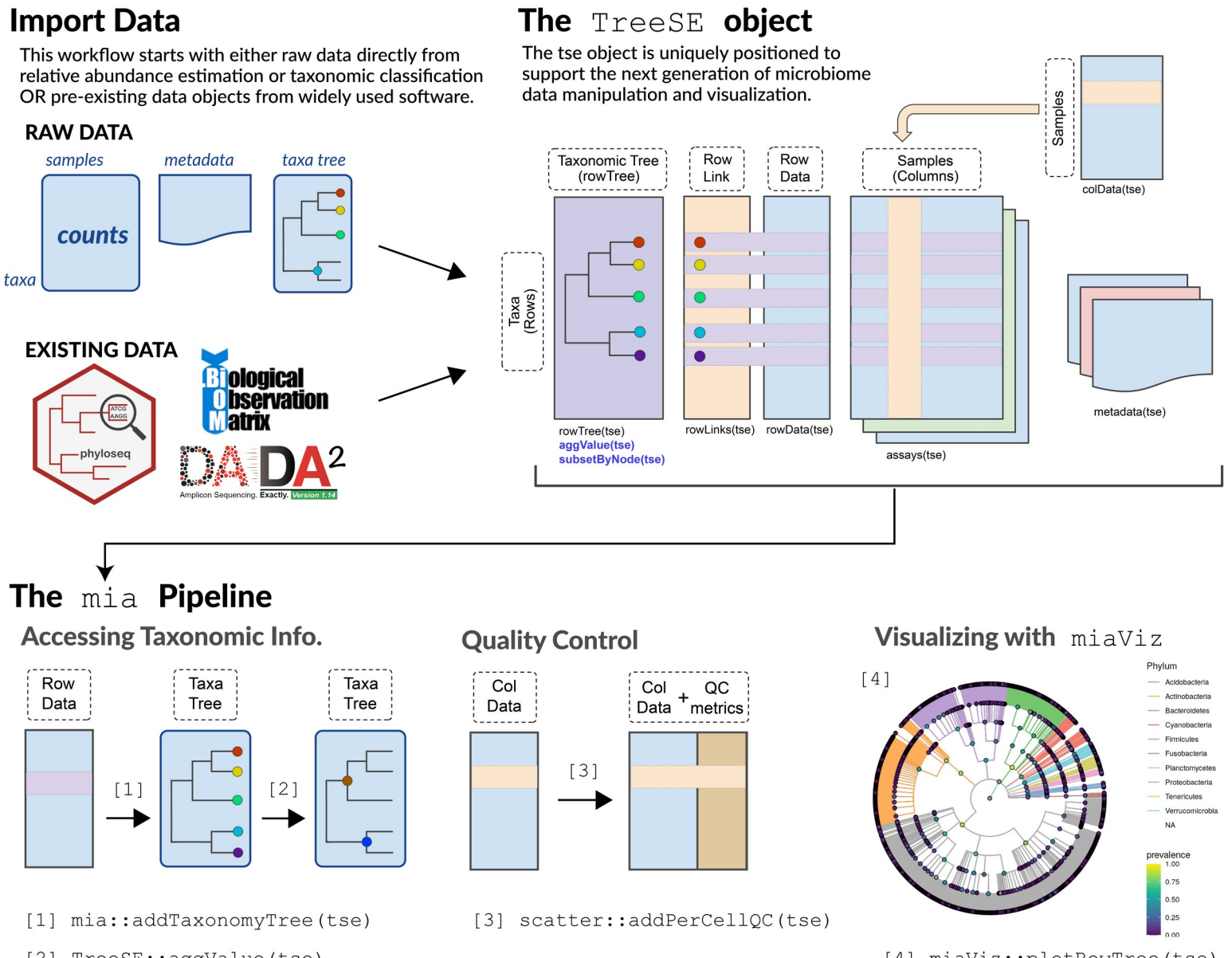
miaViz - Visualization



Package ecosystem

Example workflow – microbiome.github.io

Figure by Domenick Braccia (EuroBioC 2020)



European *Bioconductor* Meeting 2020

- Where: Virtual Conference
- When: 14-18 December 2020
- On twitter: #EuroBioc2020

What is Bioconductor ?

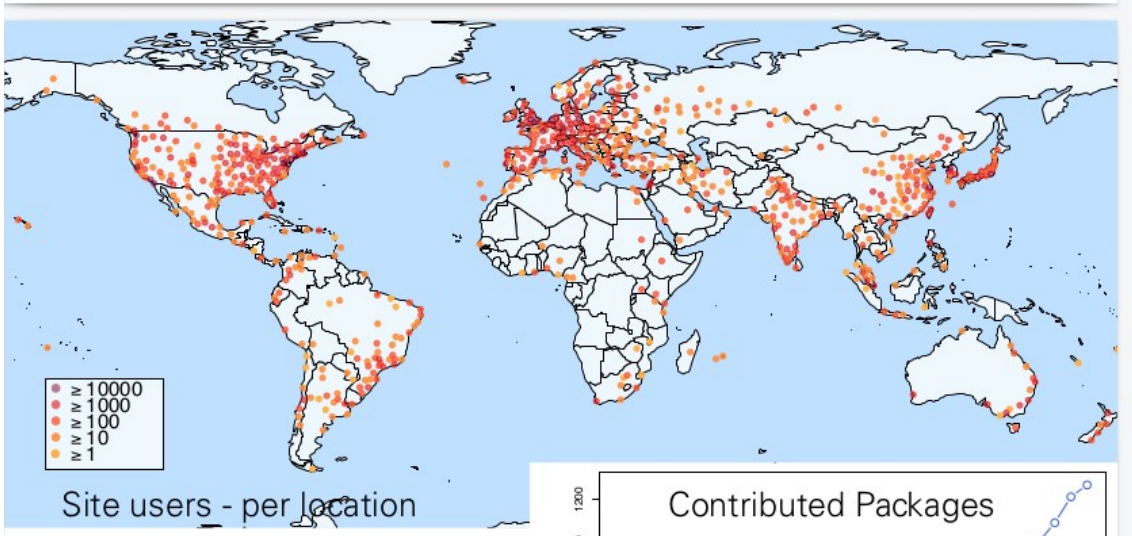
Started 2001 as a platform for analysis & understanding of microarray data

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

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

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)

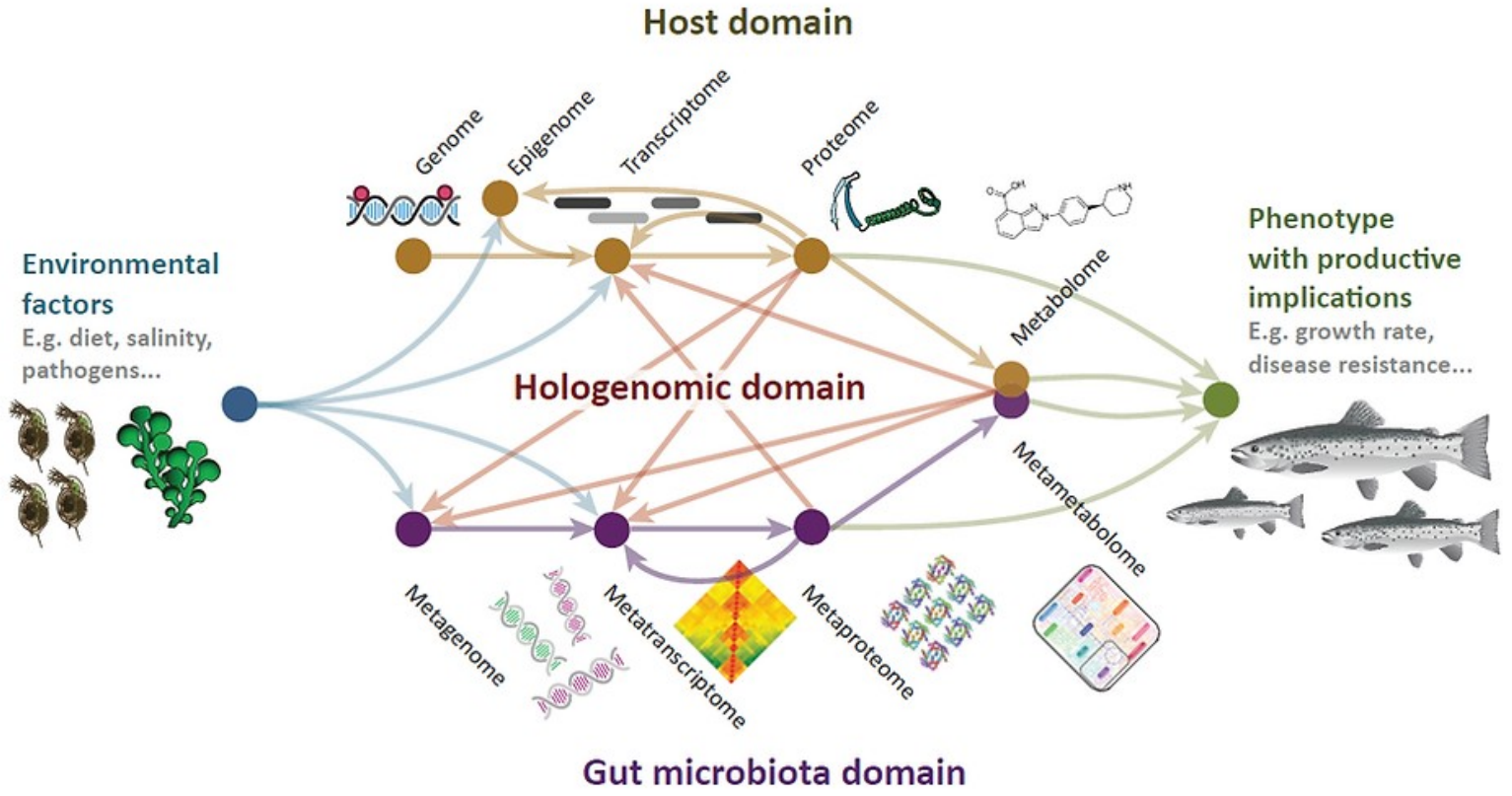
Data import and data structures

Microbiome data exploration & visualization

Playing together!

https://activeforlife.com/2020-outdoor-learning/

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.