Day 3

Morning

9-10 Lecture: multi-omic data integration

10-12 Demo & hands-on: multi-assay data container

12-13 Lunch break

Afternoon

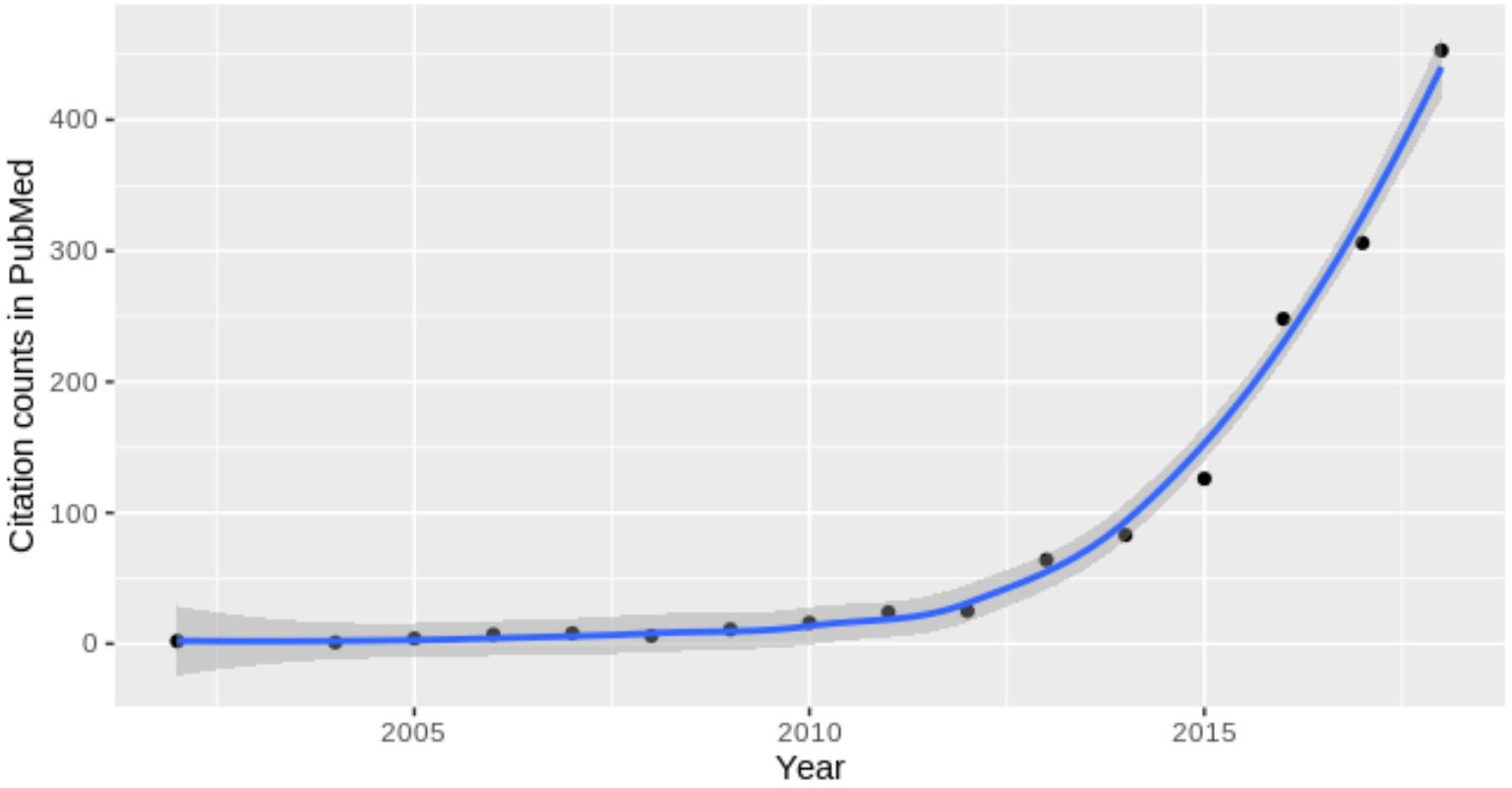
13-15 Demo & hands-on: association analysis

13-17 Demo & hands-on: machine learning

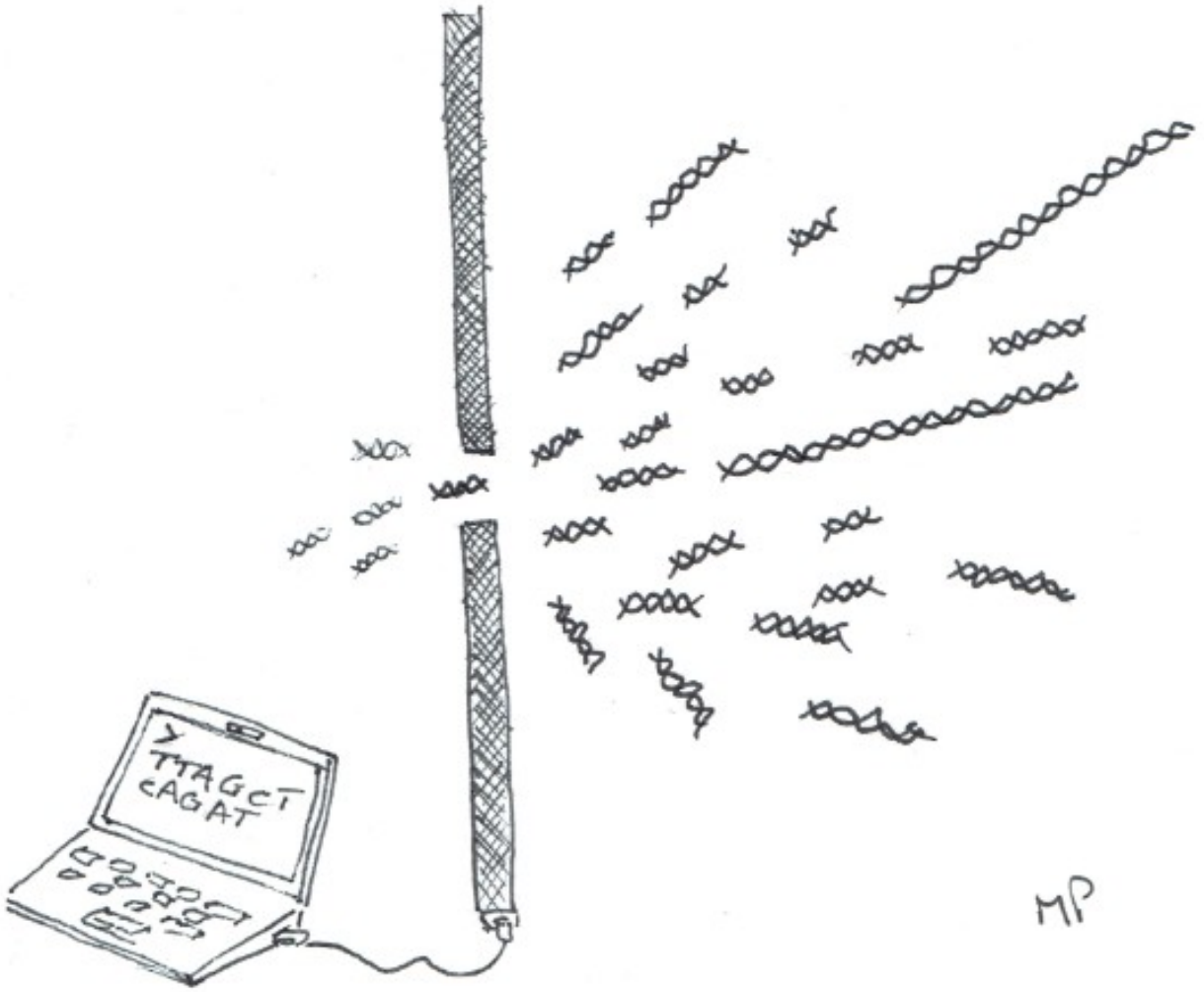
17-18 Presentations & Discussion

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

(source: Wikipedia)



Limited observations → data integration?

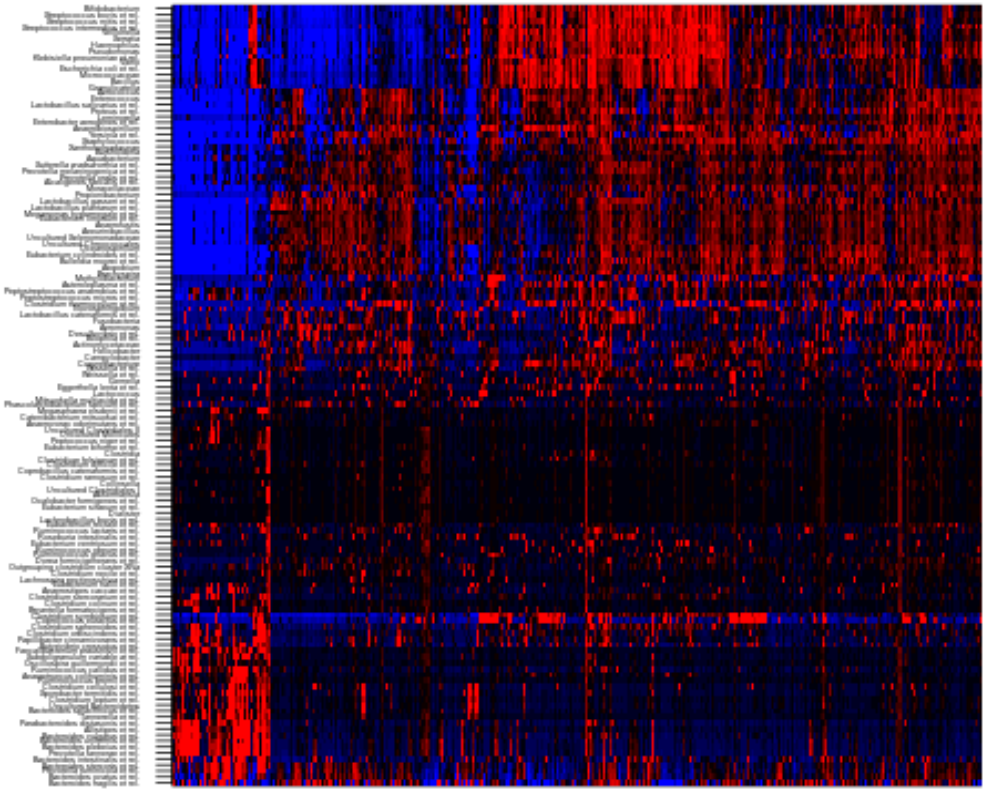


Omics:
taxonomic abundance table

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

Individuals

Taxonomic groups

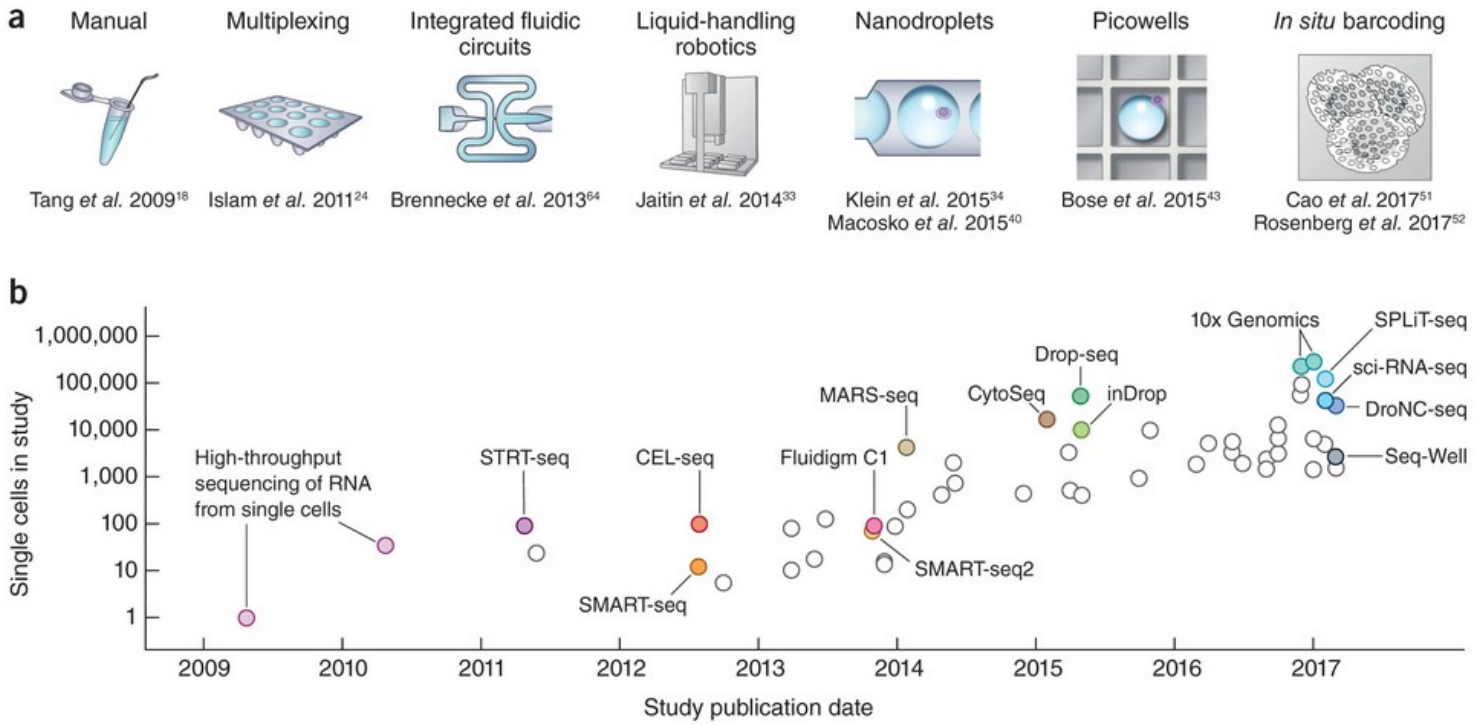


Genomics
Epigenomics
Microbiomics
Lipidomics
Proteomics
Glycomics
Foodomics
Transcriptomics
Metabolomics
Culturomics

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

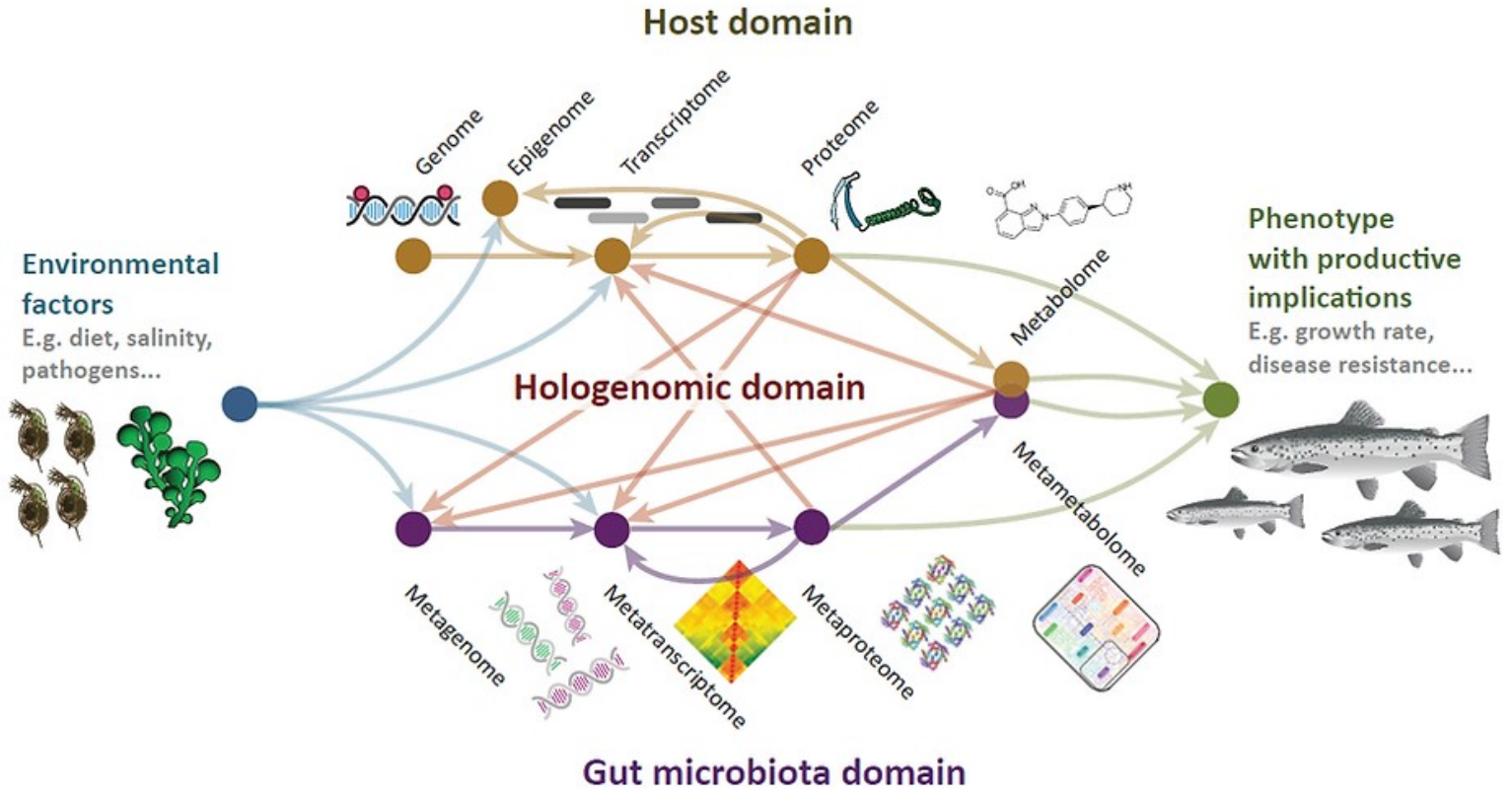
Figure 1: Scaling of scRNA-seq experiments.

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.

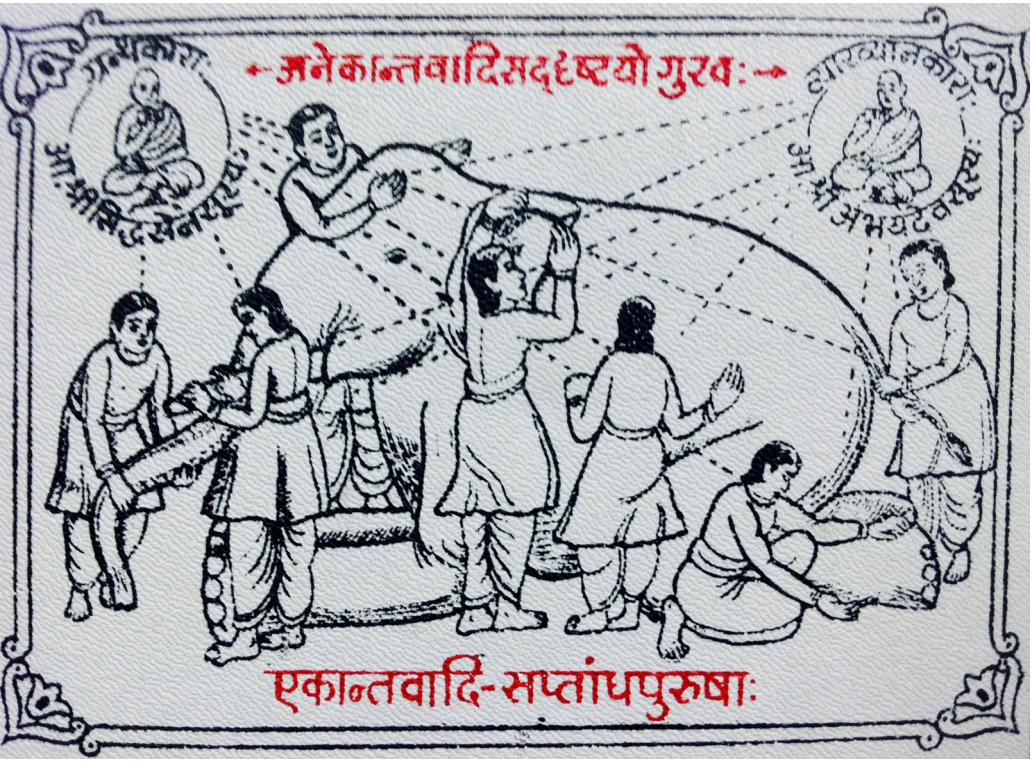
FindingPheno is creating an integrated computational framework for hologenomic big data, providing the tools to better understand how host-microbiome interactions can affect growth and other outcomes.



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.



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.

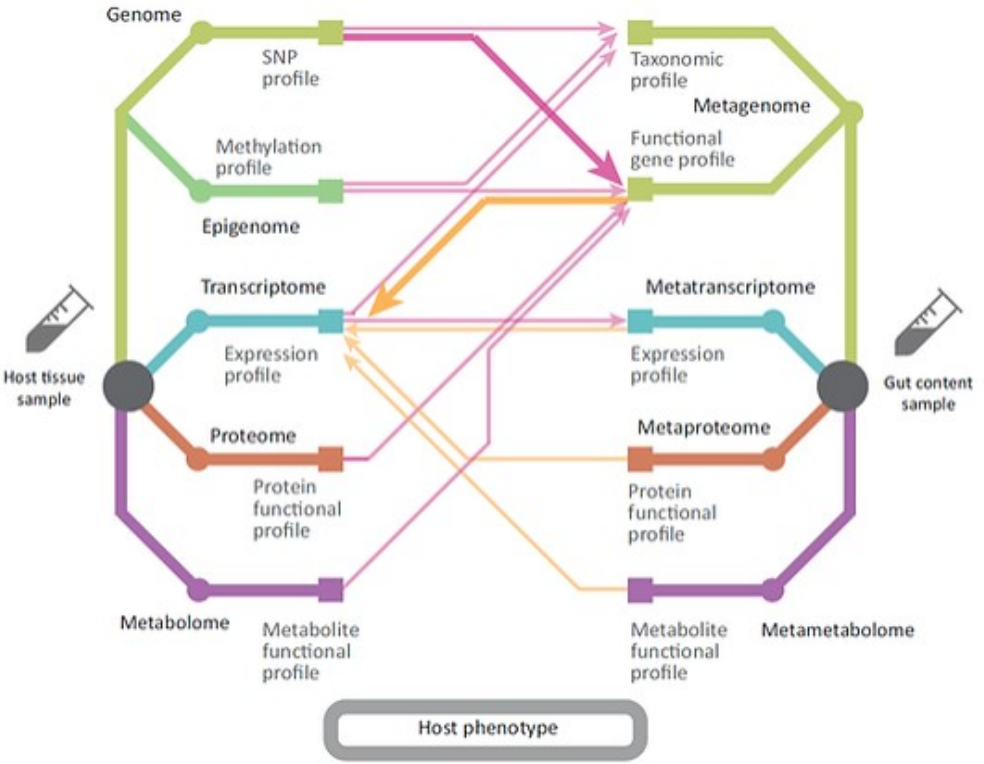
REVIEW ARTICLE
Front. Genet., 28 August 2019 | https://doi.org/10.3389/fgene.2019.00627

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.

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.

(some) topics in data integration

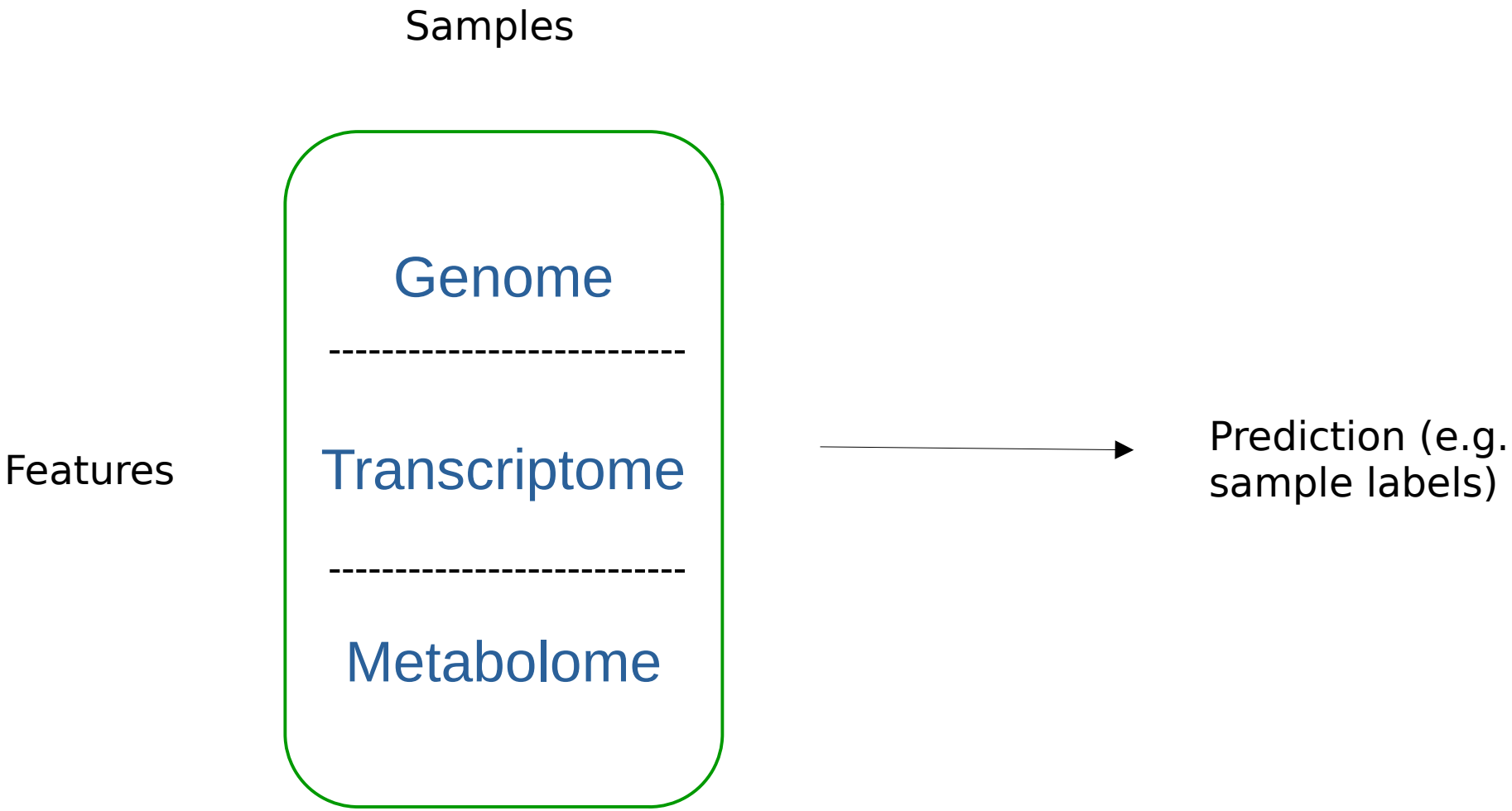
- improve predictions (of external labels)
- explore associations (btw. two or more data sets)
- identify latent processes and mechanisms
- incorporate prior knowledge

Associating data with external variables

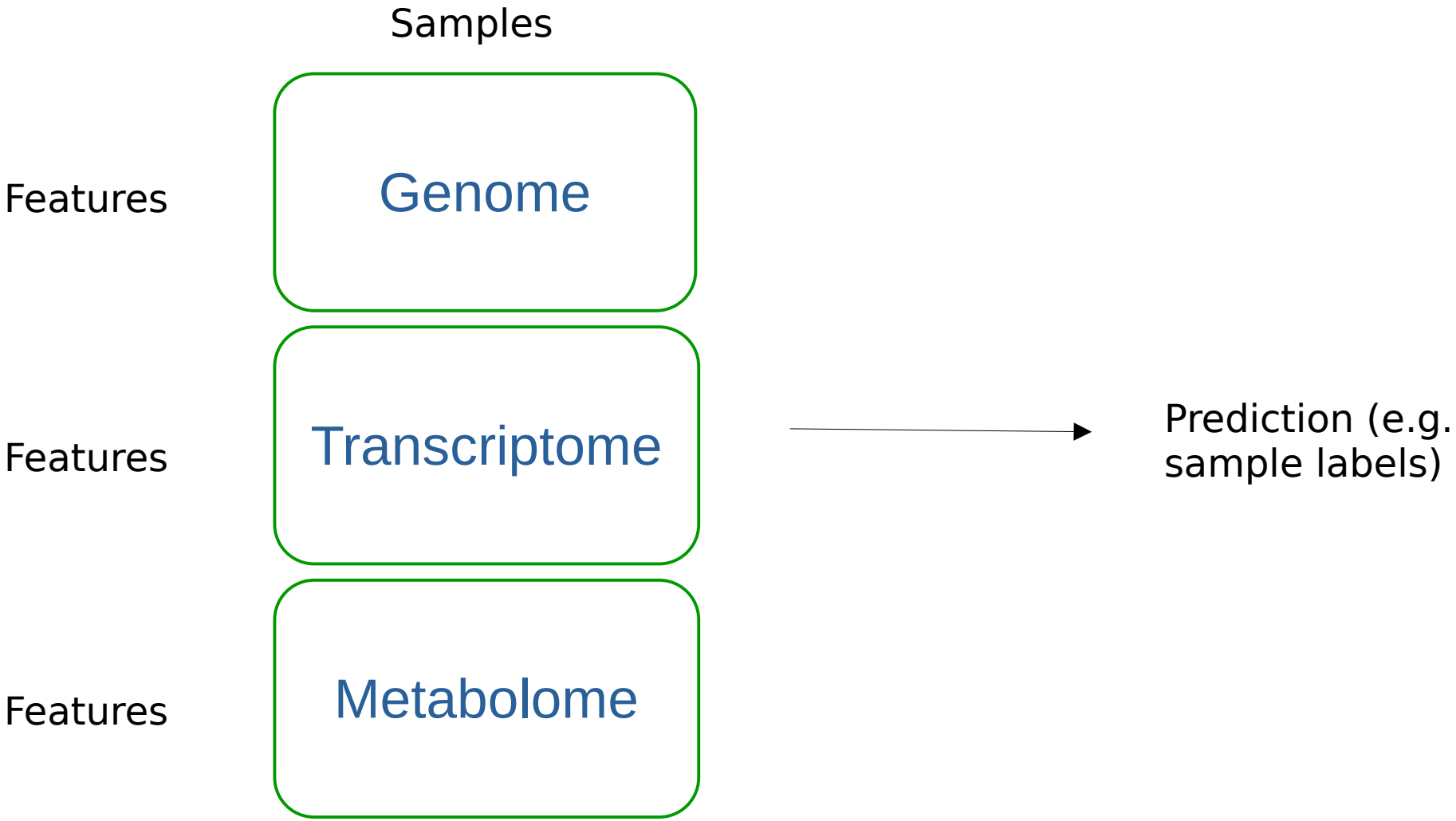
- e.g. prediction tasks

Prediction with a single data source

Concatenation: a null model for data integration?



Multi-view learning: advanced models for data integration?



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

6342 Accesses | **38** Citations | **4** Altmetric | Metrics

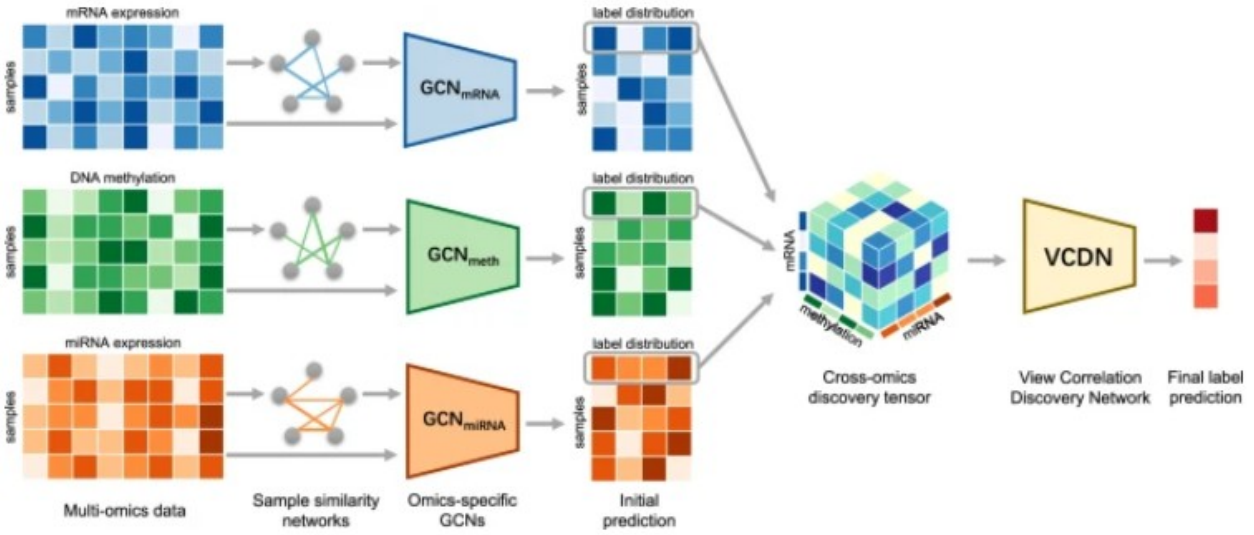
MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification

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

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

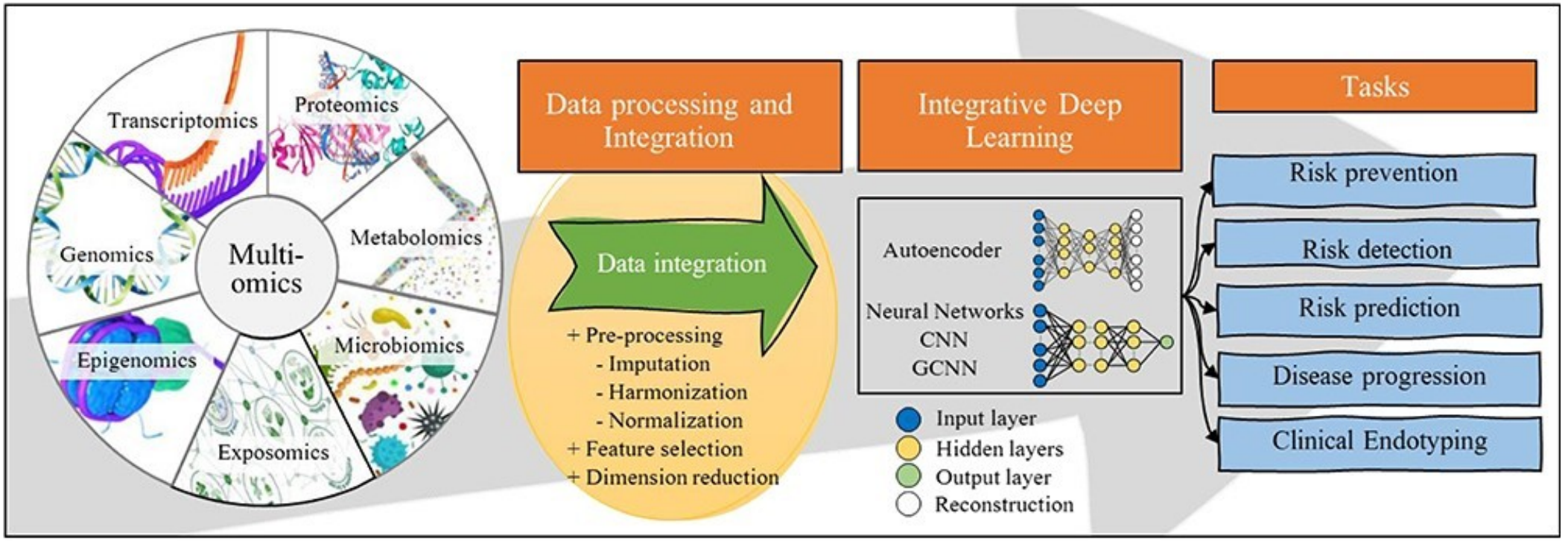
7874 Accesses | **3** Citations | **40** Altmetric | Metrics

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

“Interpretable” neural network

Computational and Structural Biotechnology Journal
Volume 19, 2021, Pages 3735-3746

Integration strategies of multi-omics data for machine learning analysis

Milan Picard [a], Marie-Pier Scott-Boyer [a], Antoine Bodein [a], Olivier Périn [b], Arnaud Droit [a]

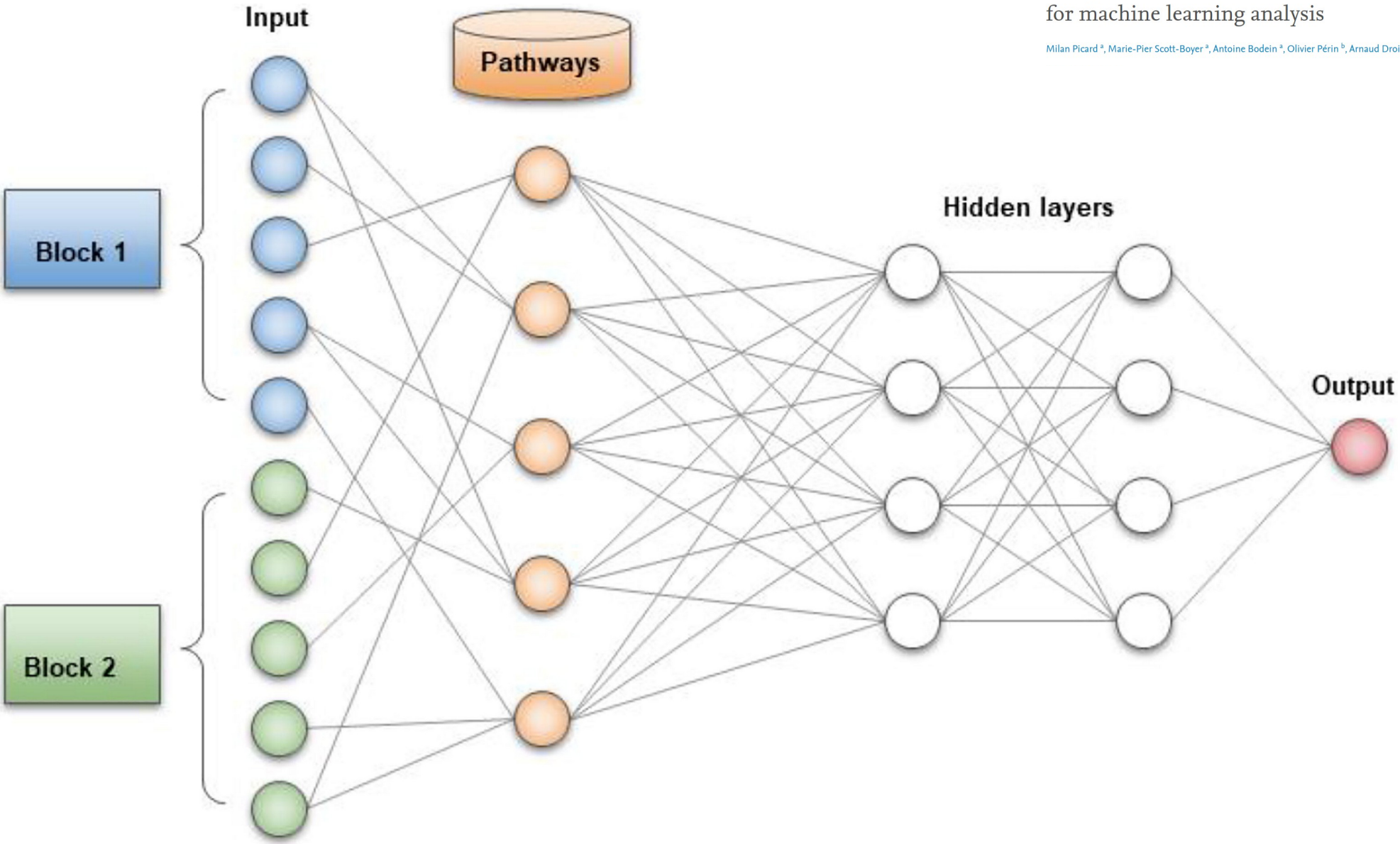


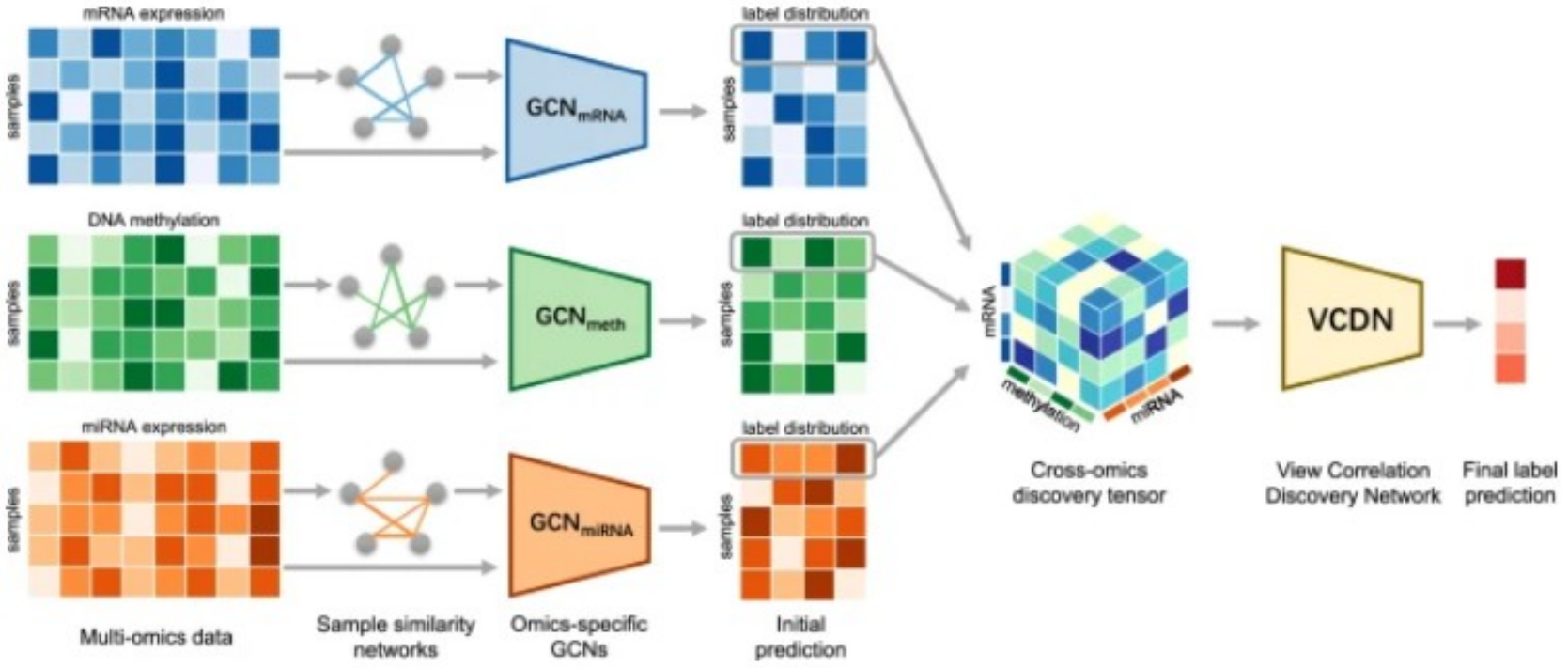
Fig. 1. Structure of an interpretable artificial neural network. The input layer is followed by an additional pathway layer, where each node corresponds to a known molecular pathway. If a molecule is known to be involved in a pathway, a connection is made between the two. Hence, important pathways implicated in the outcome are activated with bigger weights during training. Figure inspired from Deng et al. (2020).

Improving predictions by data integration

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

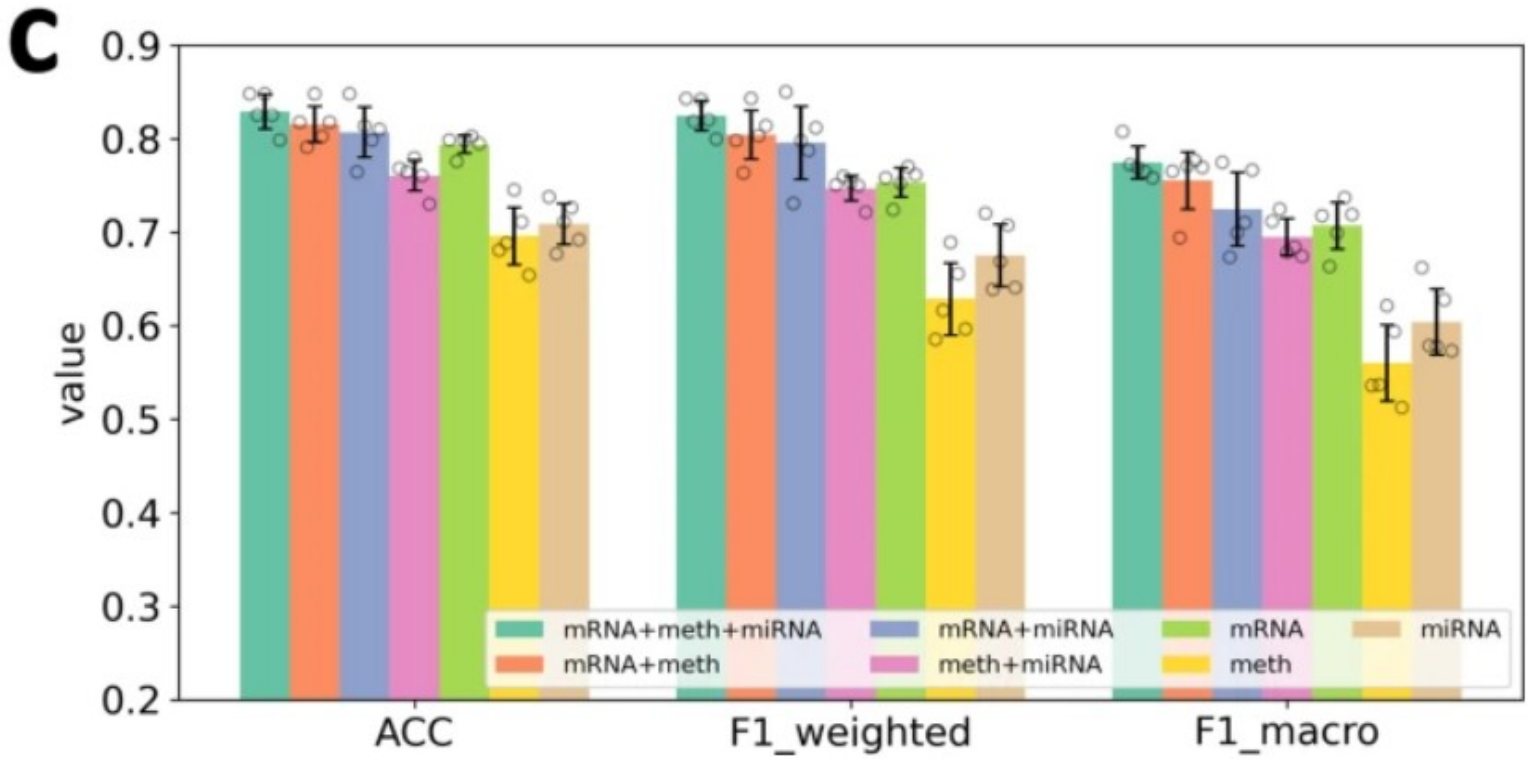
Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding ✉ & Kun Huang ✉

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



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.

Data integration leads to improved predictions in a BRCA data set



a Results of the ROSMAP dataset. **b** Results of the LGG dataset. **c** Results of the BRCA dataset. Means of evaluation metrics with standard deviations from different experiments are shown in the figure, where the error bar represents plus/minus one standard deviation. mRNA, meth, and miRNA refer to single-omics data classification via GCN with mRNA expression data, DNA methylation data, and miRNA expression data, respectively. mRNA + meth, mRNA + miRNA, and meth + miRNA refer to classification with two types of omics data. mRNA + meth + miRNA refers to classification with three types of omics data. Source data are provided as a Source Data file.

A non-exhaustive list of multi-block dimensionality reduction methods for multi-omics datasets. NMF: Non-negative Matrix Factorization, MOFA: Multi-Omics Factor Analysis, JIVE: Joint and Individual Variation Explained, MO: multi-omic.

Method	Principle	Purpose	Recent applications
jNMF/intNMF/nNMF [132], [133], [139]	Matrix factorization	Disease subtyping, module detection, biomarker discovery	jNMF found biomarkers in MO and pharmacological data connected to drug sensitivity in cancerous cell lines [140]. intNMF identified Glioblastoma and breast cancer subtypes from MO and clinical data [134].
MOFA/MOFA+ [141], [142]	Bayesian Factor Analysis	biomarker discovery, systemic knowledge	MOFA found new biomarkers and pathways associated with Alzeihmer's disease based on MO data including proteomics, metabolomics, lipidomics [143].MOFA + found predictive biomarkers from DNA methylation and gene expression data in cardiovascular disease [144].
iCluster [145]	Gaussian latent variable model Generalized linear regression Bayesian integrative clustering	Disease subtyping, biomarker discovery	iCluster was used to identify subtypes of esophageal carcinoma from genomic, epigenomic and transcriptomic data [148].
iClusterPlus [146]			iClusterPlus was used to identify subtypes of non-responsive samples with ovarian cancer from different omics datasets [149].
iClusterBayes [147]			iClusterBayes was used to identify predictive biomarkers and clinically relevant subtypes on MIB cancer from 5 different omics [150].
JIVE/aJIVE [151], [152]	Matrix factorization	Disease subtyping, systemic knowledge, module detection	JIVE was used as a dimension reduction technique to improve survival prediction of patients with glioblastoma from mRNA, miRNA and methylation data [153].
Integrated PCA [64]	Generalized PCA	Visualization, prediction	iPCA was used as a dimension reduction technique to improve prediction of outcome on lung cancer from CpG methylation data, mRNA and miRNA expression [154].
SLIDE [130]	Matrix factorization	Disease subtyping, module detection, biomarker discovery	SLIDE was used on DNA methylation data and gene, protein and miRNA expression for subtyping patients with breast cancer [130].

ELSEVIER

Computational and Structural Biotechnology Journal

Volume 19, 2021, Pages 3735-3746

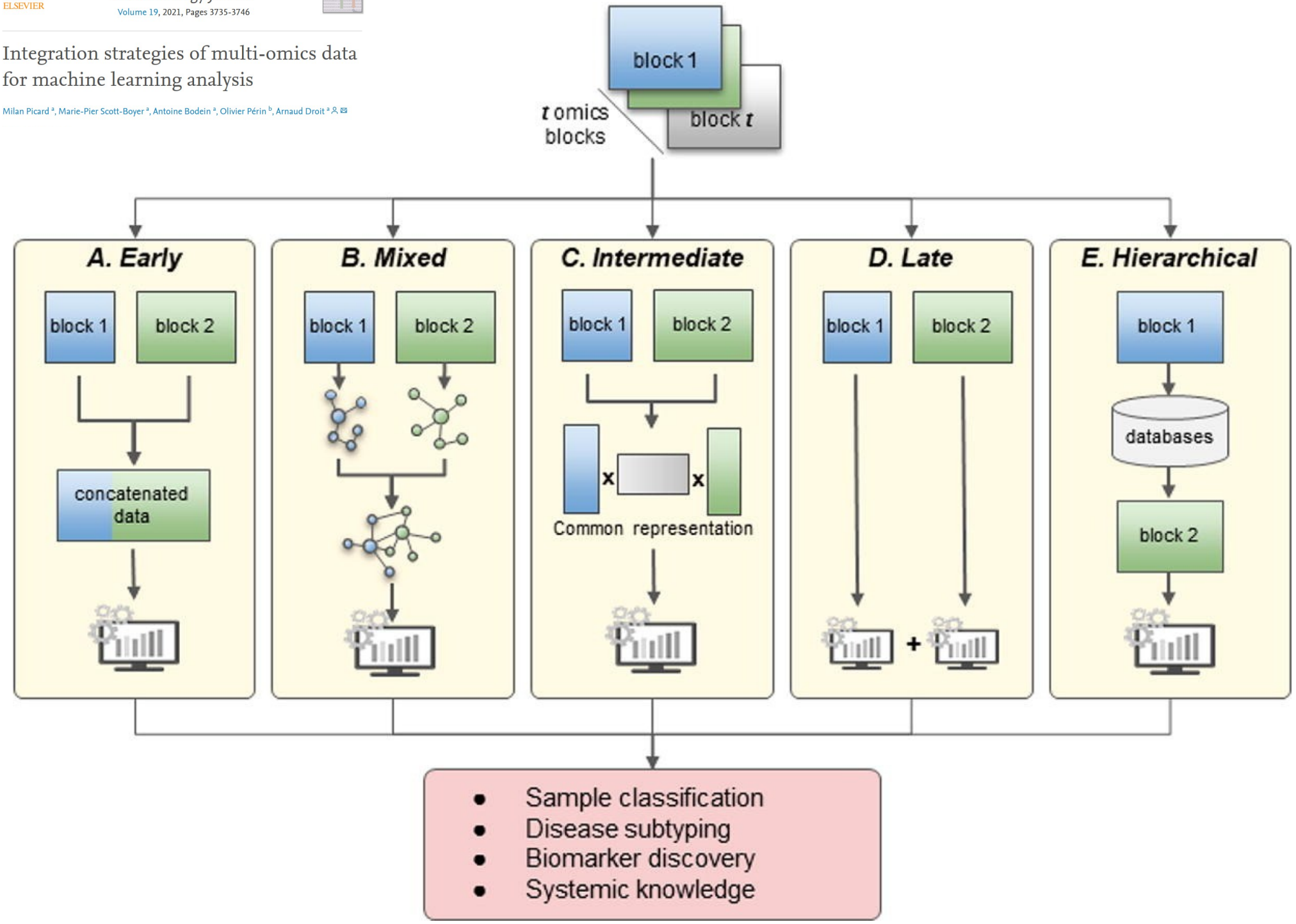
Integration strategies of multi-omics data for machine learning analysis

Milan Picard [a], Marie-Pier Scott-Boyer [a], Antoine Bodein [a], Olivier Périn [b], Arnaud Droit [a]

Computational and Structural Biotechnology Journal
Volume 19, 2021, Pages 3735-3746

Integration strategies of multi-omics data for machine learning analysis

Milan Picard [a], Marie-Pier Scott-Boyer [a], Antoine Bodein [a], Olivier Périn [b], Arnaud Droit [a]



(some) questions in data integration

- **a/symmetry** between data sets?

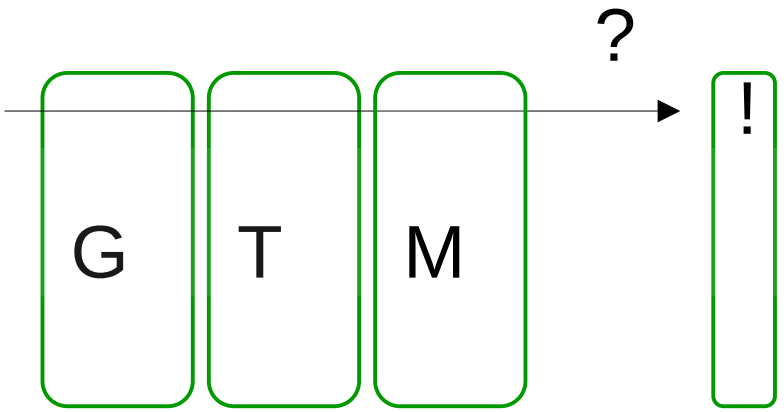
- **scale**: small mechanistic models vs. large-scale exploration?

- **two or more data sets**?

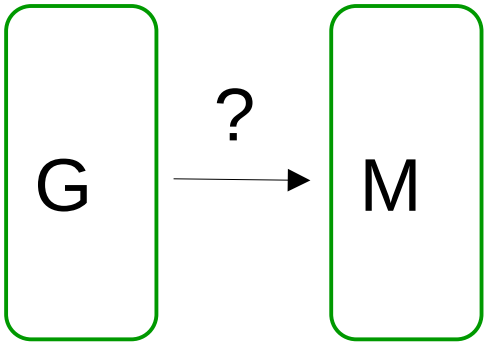
- **known structure**?

- **computational requirements**?

Predicting external labels

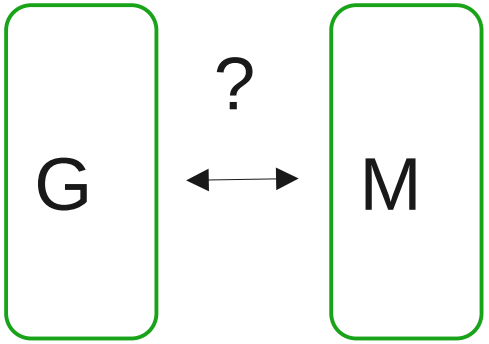


Association -
one data set is primary



Association -

all data sets are equal



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

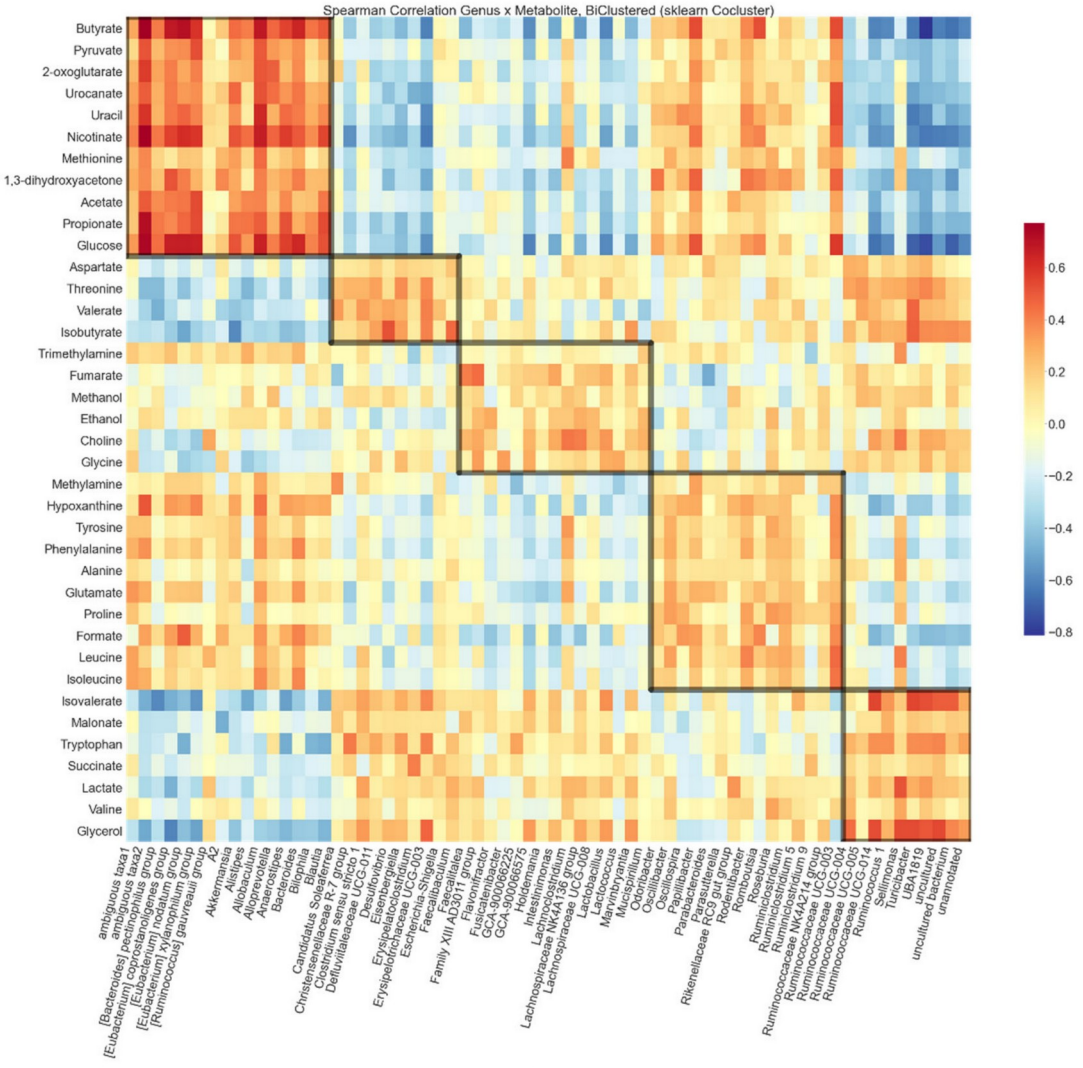
Open Access Article

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,†]

Simple approach:
- cross-correlate data sets
- visualize
- characterize

How to generalize to more than two data sets..?



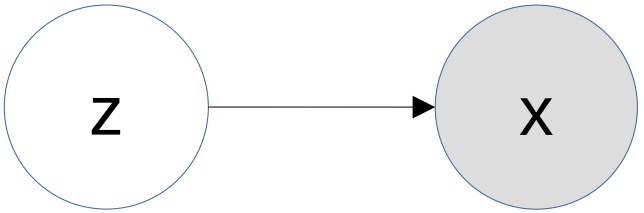
(a)

Example: PCA vs. CCA

PCA: Principal component analysis
→ captures *maximal variation* in a single data set

CCA: Canonical correlation analysis
→ captures *maximal correlation* between two data sets

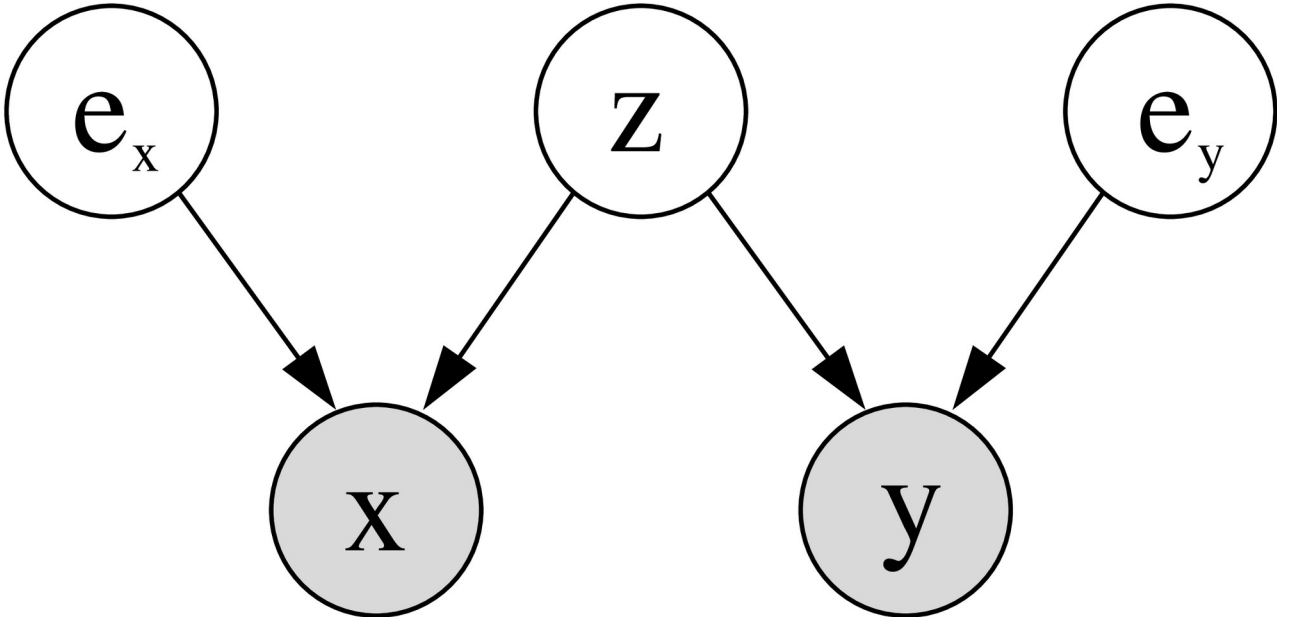
Probabilistic PCA

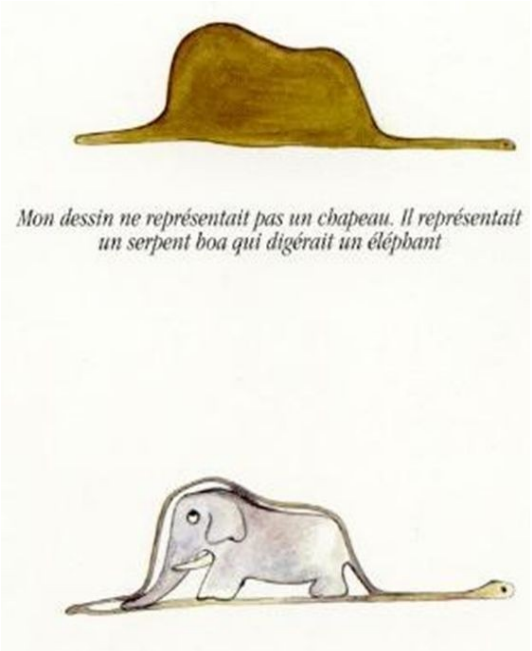


$$X = W_x \mathbf{z} + \varepsilon_x$$

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

Multi-view learning





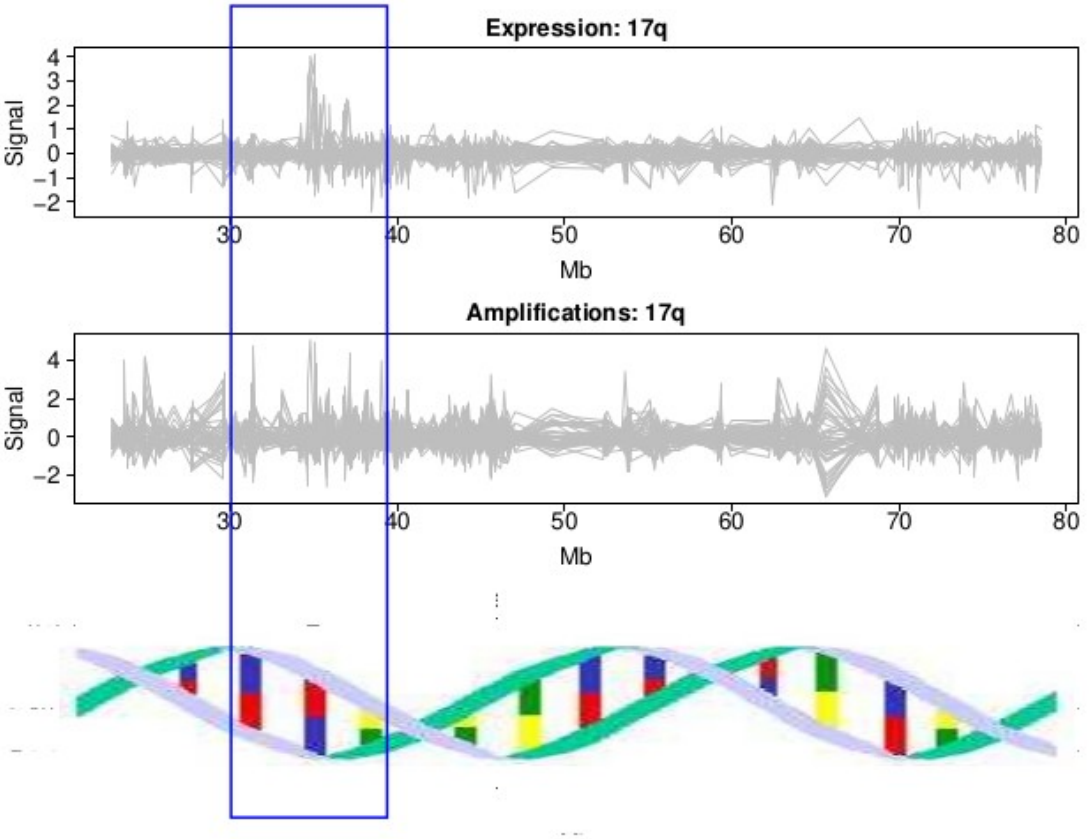
https://github.com/mblstamps/stamps2019/blob/master/STAMPS2019_overview_Pop.pdf

Chromosome arm 17q

Investigate dependencies within local chromosomal regions using sliding window

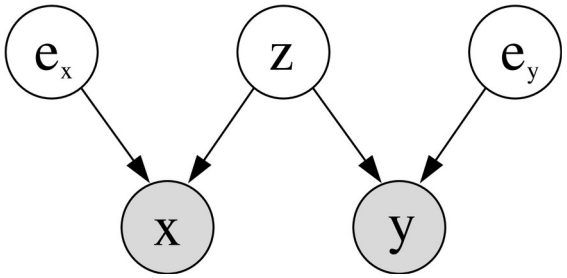
X

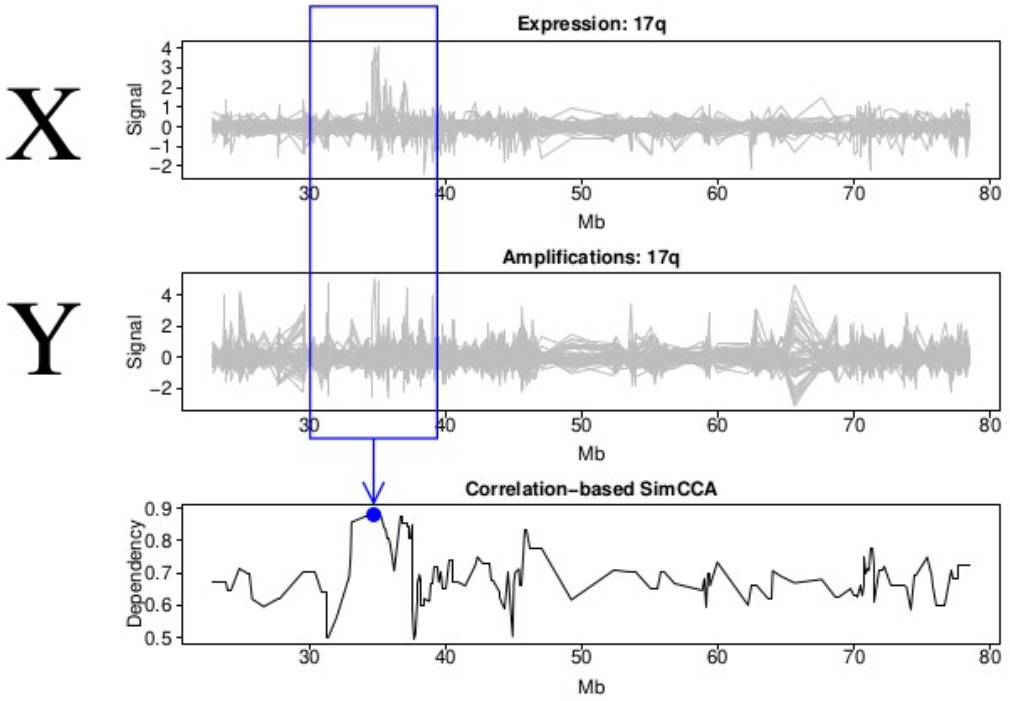
Y



Chromosome arm 17q: results

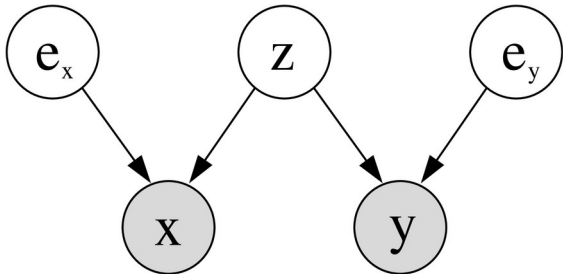
SimCCA measures dependency between data sources within each chromosomal region

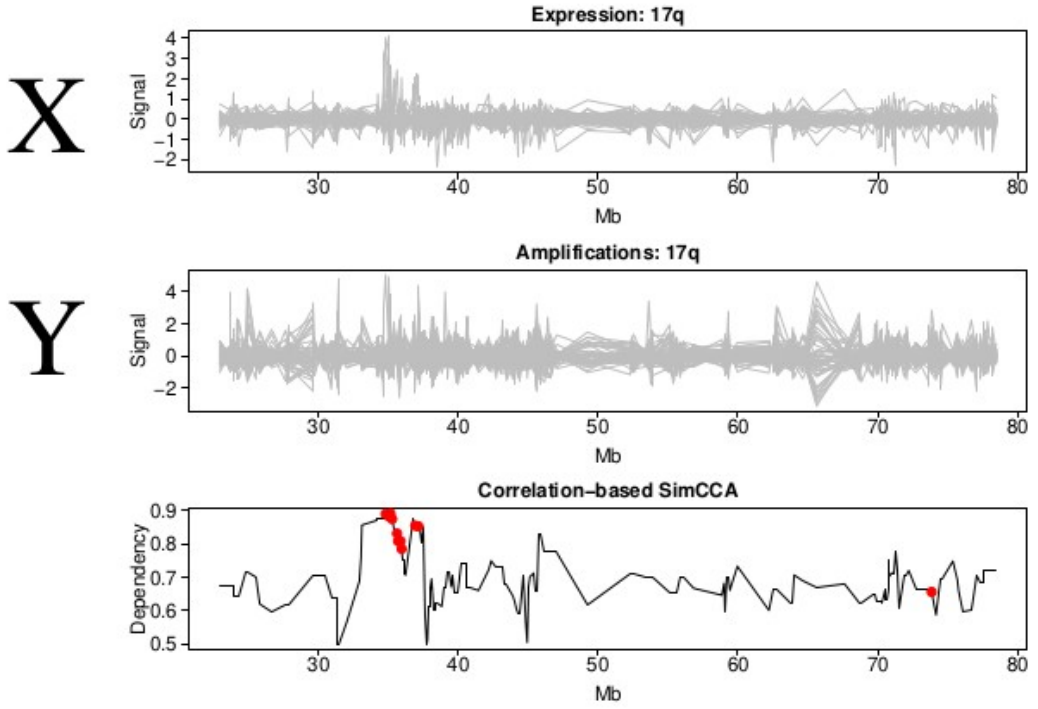




Chromosome arm 17q: results

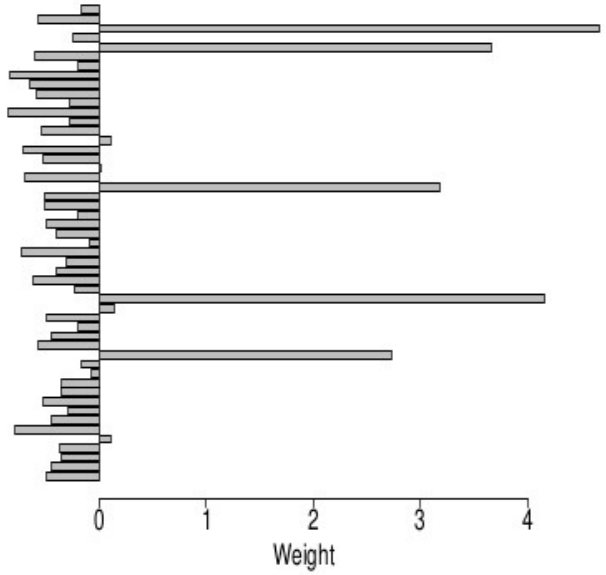
SimCCA reveals known gastric cancer-associated chromosomal regions



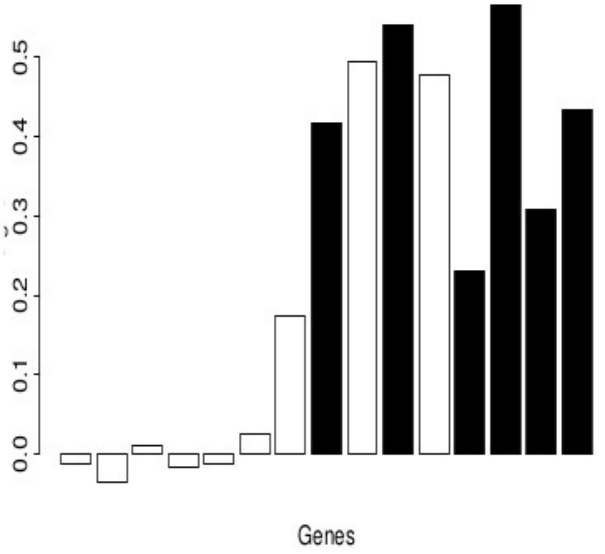


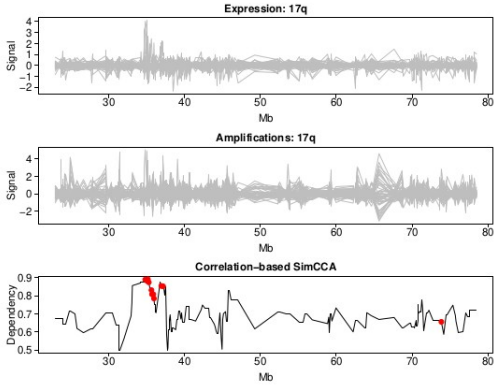
Interpreting the parameters

Z: affected patients

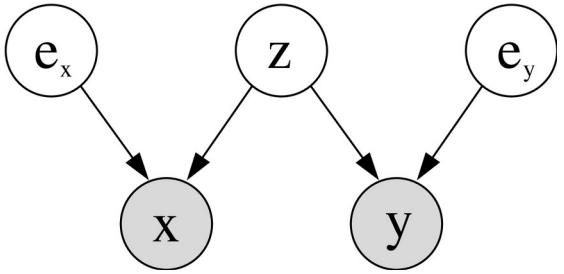


W: dependent observations





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



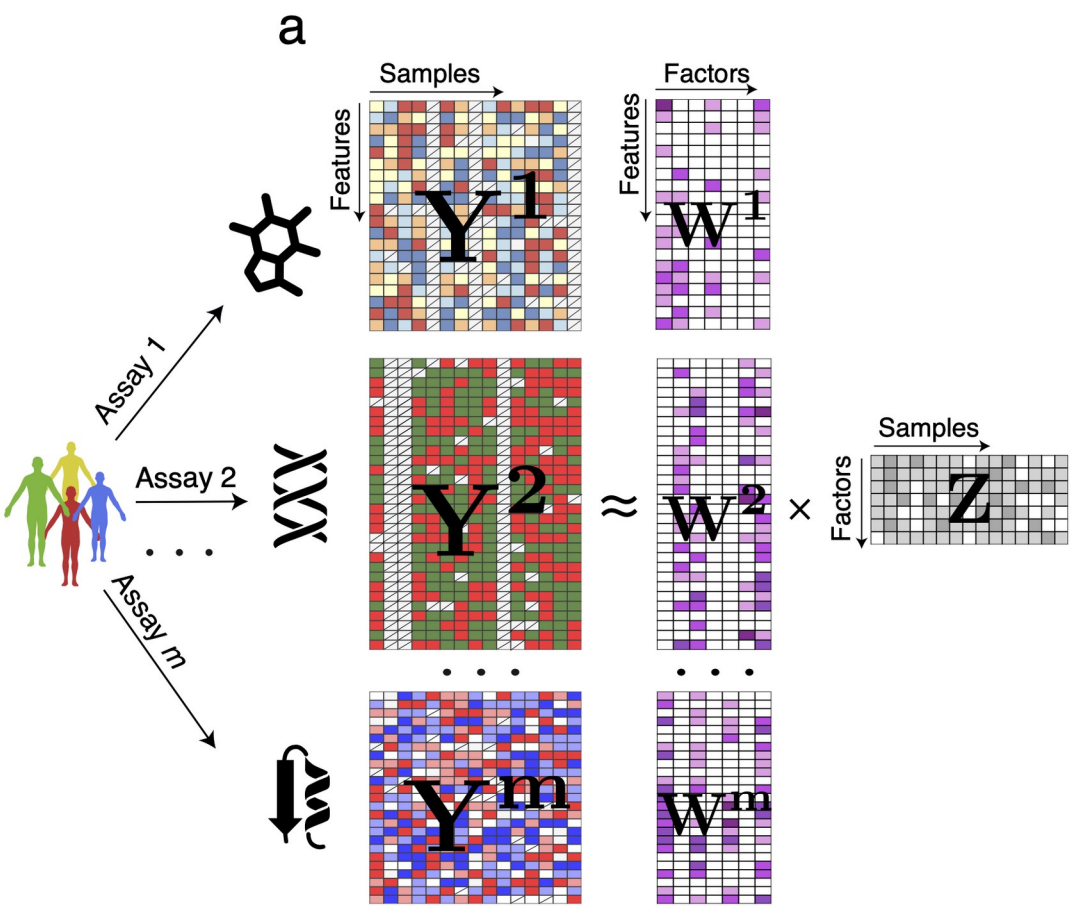
Method | 20 June 2018 | OPEN ACCESS TRANSPARENT PROCESS

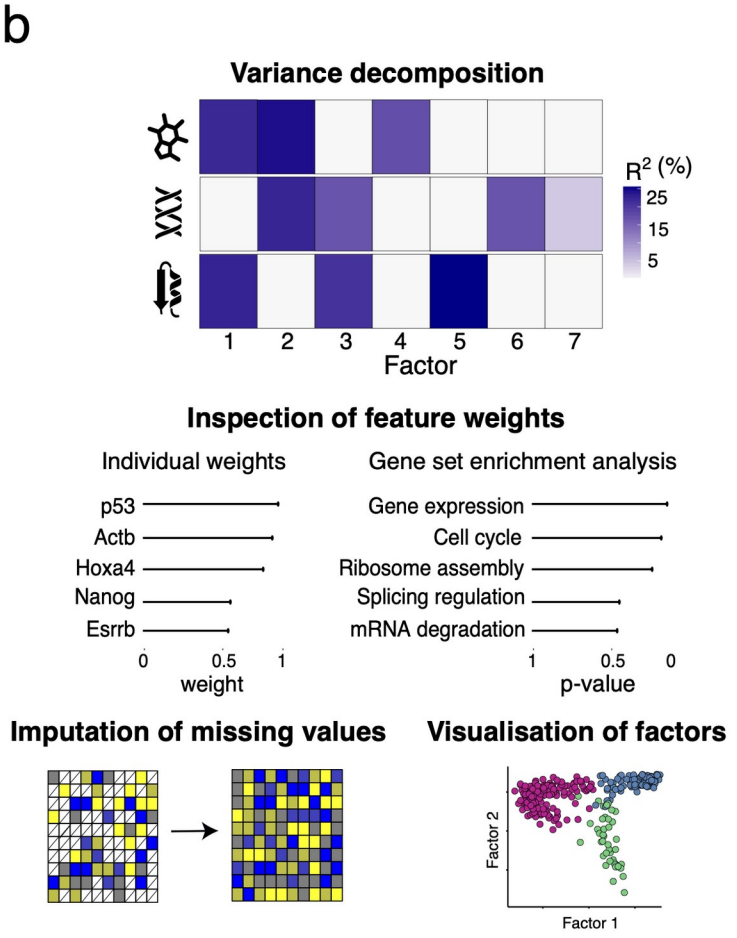
Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets

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/

Method | Open Access | Published: 11 May 2020

MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data

Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni & Oliver Stegle

Genome Biology **21**, Article number: 111 (2020) | Cite this article

From: MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data

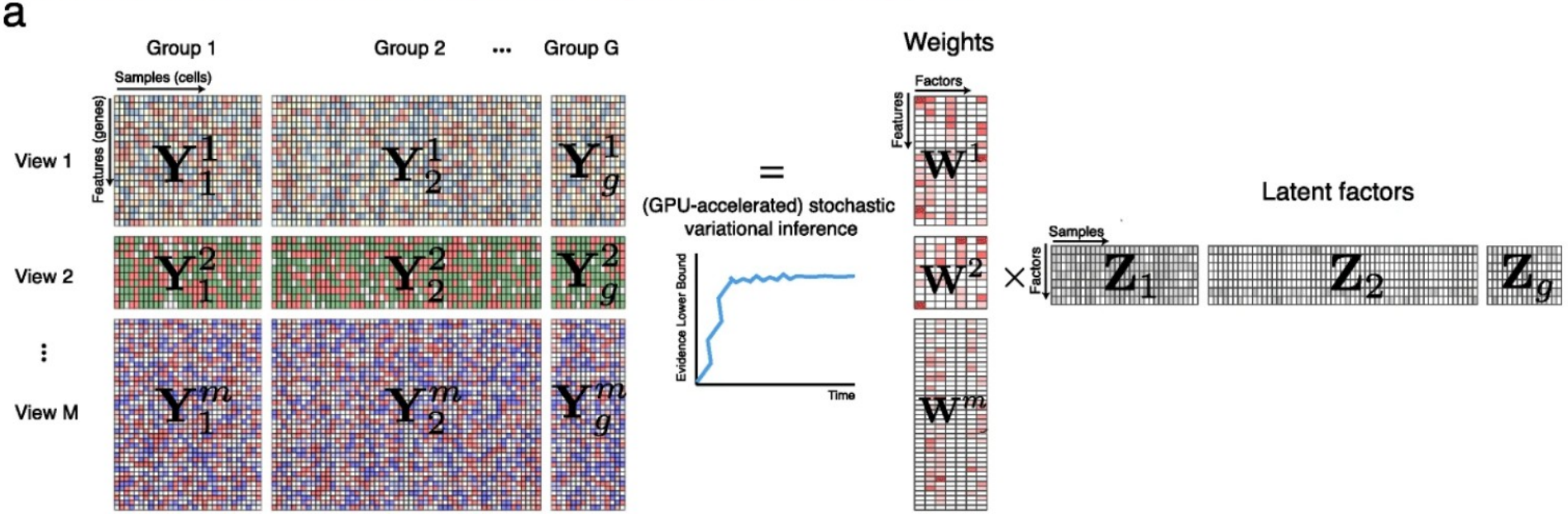
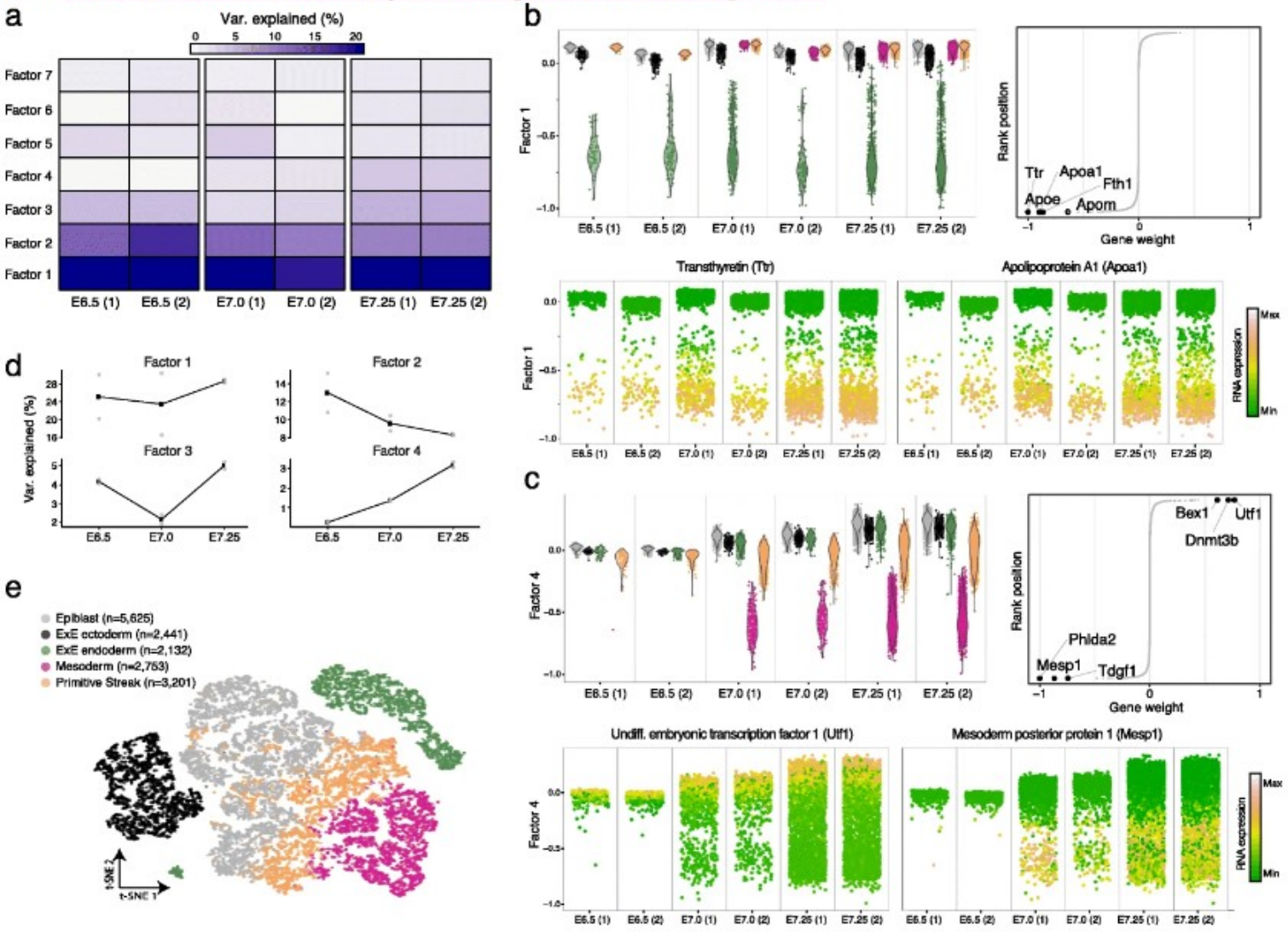


Fig. 2

From: MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data



Integration of heterogeneous scRNA-seq experiments reveals stage-specific transcriptomic signatures associated with cell type commitment in mammalian development. **a** The heatmap displays the percentage of variance explained for each Factor (rows) in each group (pool of mouse embryos at a specific developmental stage, columns). **b**, **c** Characterization of Factor 1 as extra-embryonic (ExE) endoderm formation (**b**) and Factor 4 as Mesoderm commitment (**c**). In each panel, the top left plot shows the distribution of Factor values for each batch of embryos. Cells are colored by cell type. Line plots (top right) show the distribution of gene weights, with the top five genes with largest (absolute) weight highlighted. The bottom beeswarm plots represent the distribution of Factor values, with cells colored by the expression of the genes with highest weight. **d** Line plots show the percentage of variance explained (averaged across the two biological replicates) for each Factor as a function of time. The value of each replicate is shown as gray dots. **e** Dimensionality reduction using t-SNE on the inferred factors. Cells are colored by cell type

Method | Open Access | Published: 11 May 2020

MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data

Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni & Oliver Stegle

Genome Biology **21**, Article number: 111 (2020) | Cite this article

18k Accesses | **54** Citations | **123** Altmetric | Metrics

Online learning / prior information

Article | Published: 19 April 2021

Iterative single-cell multi-omic integration using online learning

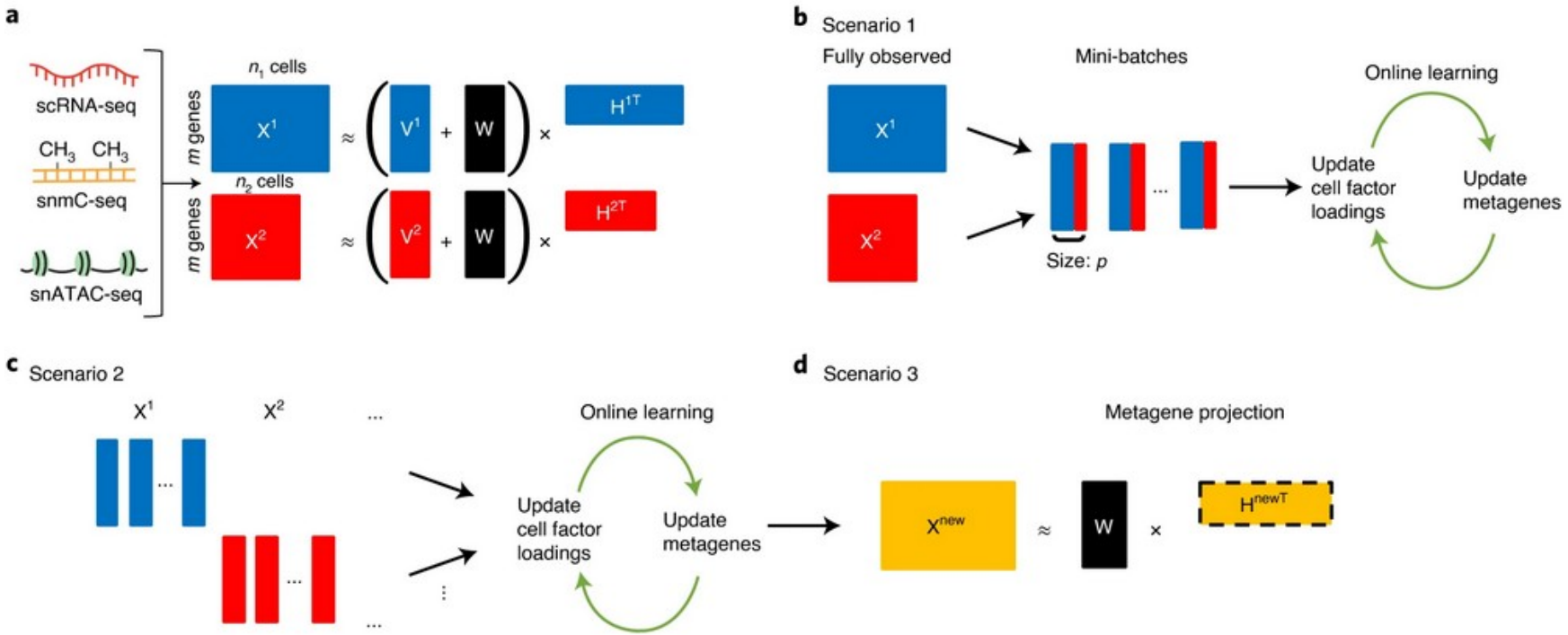
Chao Gao, Jialin Liu, April R. Kriebel, Sebastian Preissl, Chongyuan Luo, Rosa Castanon, Justin Sandoval, Angeline Rivkin, Joseph R. Nery, Margarita M. Behrens, Joseph R. Ecker, Bing Ren & Joshua D. Welch

Nature Biotechnology **39**, 1000–1007 (2021) | Cite this article

10k Accesses | **4** Citations | **130** Altmetric | Metrics

Fig. 1: Overview of the online iNMF algorithm.

From: Iterative single-cell multi-omic integration using online learning



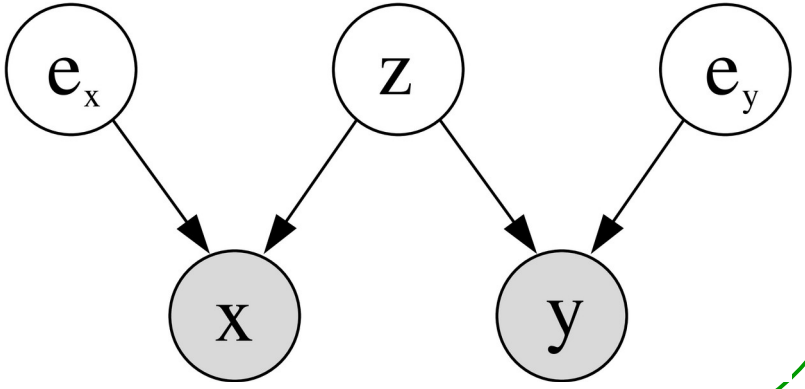
a, Schematic of iNMF: the input single-cell datasets are jointly decomposed into shared (W) and dataset-specific (V^i) metagenes and corresponding 'metagene expression levels' or cell factor loadings (H^i). These metagenes and cell factor loadings provide a quantitative definition of cell identity and how it varies across biological settings. **b**–**d**, Three different scenarios in which online learning can be used for single-cell data integration. **b**, Scenario 1: the single-cell datasets are large but fully observed. Online iNMF processes the data in random mini-batches, enabling memory usage independent of dataset size. Each cell may be used multiple times in different epochs of training to update the metagenes. **c**, Scenario 2: the datasets arrive sequentially, and online iNMF processes the datasets as they arrive, using each cell to update the metagenes exactly once. **d**, Scenario 3: online iNMF is performed as in Scenario 1 or Scenario 2 to learn W and V^i. Then cell factor loadings for the newly arriving dataset are calculated using the shared metagenes (W) learned from previously processed datasets. The new dataset is not used to update the metagenes.

Take-home messages

- Heterogeneity of problems

- Role of bias & noise, need for data-specific customization

- Importance of study question



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

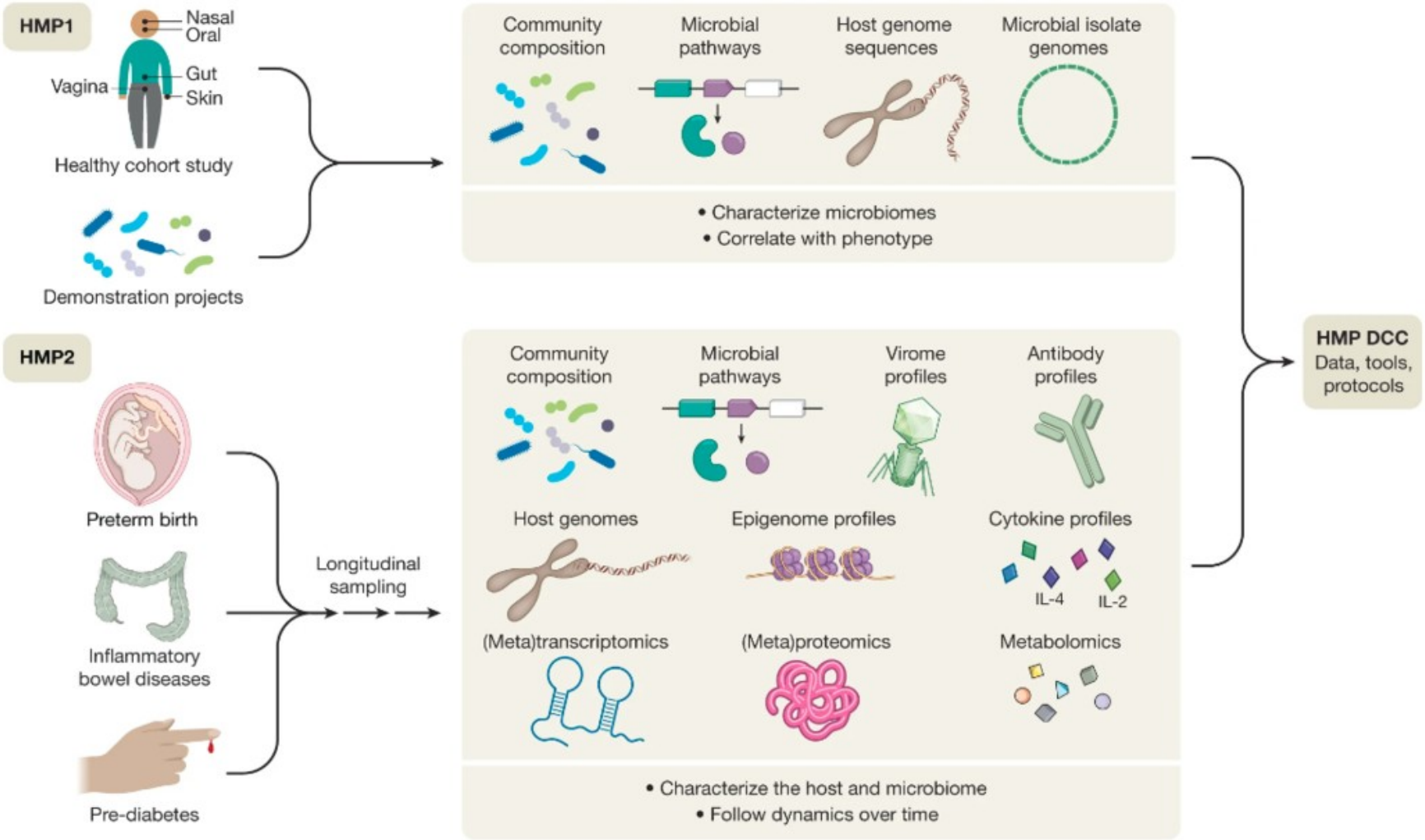
Authors | Authors and affiliations

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

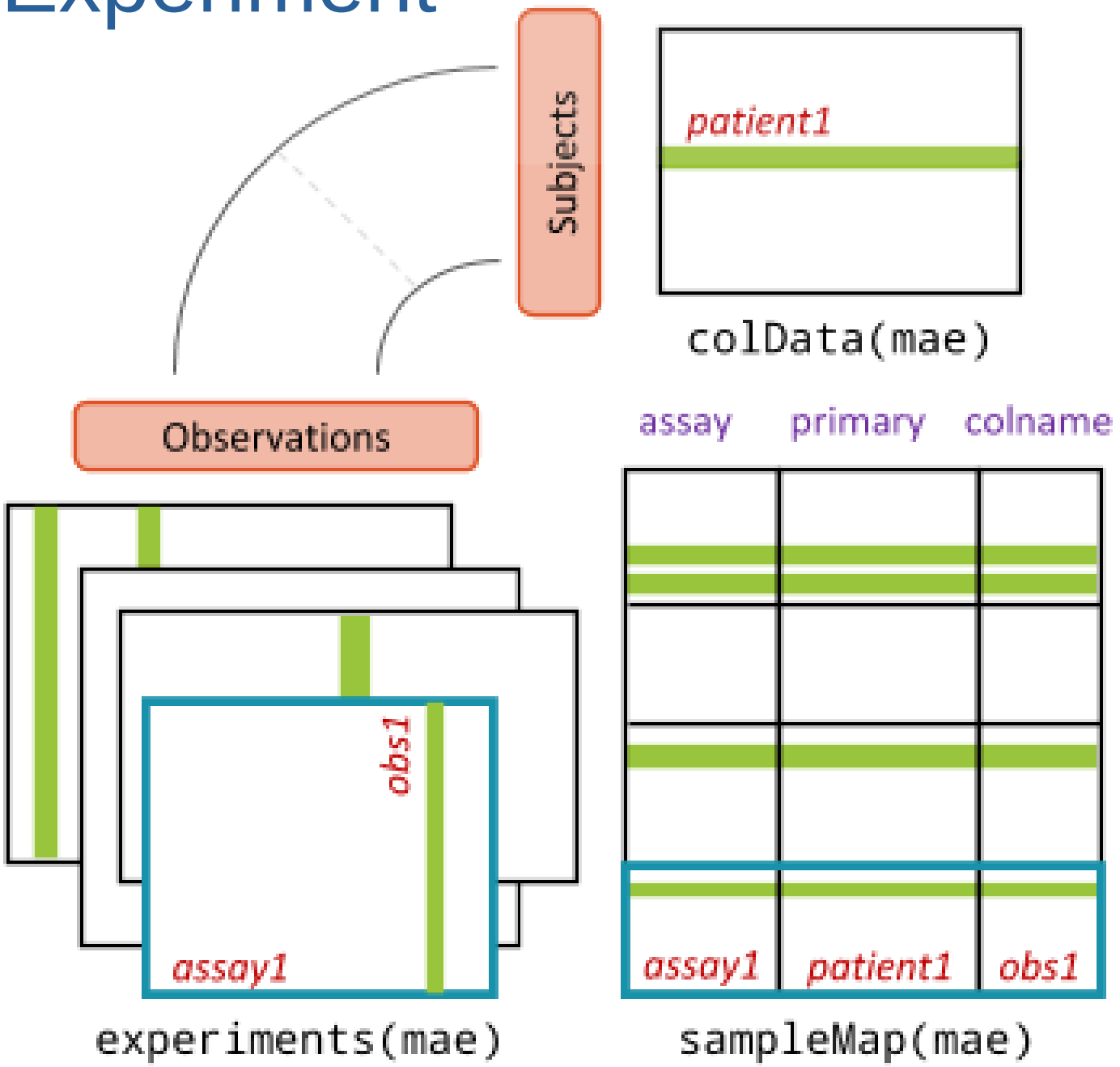
→ Compatibility?

Multi-omics



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

MultiAssayExperiment

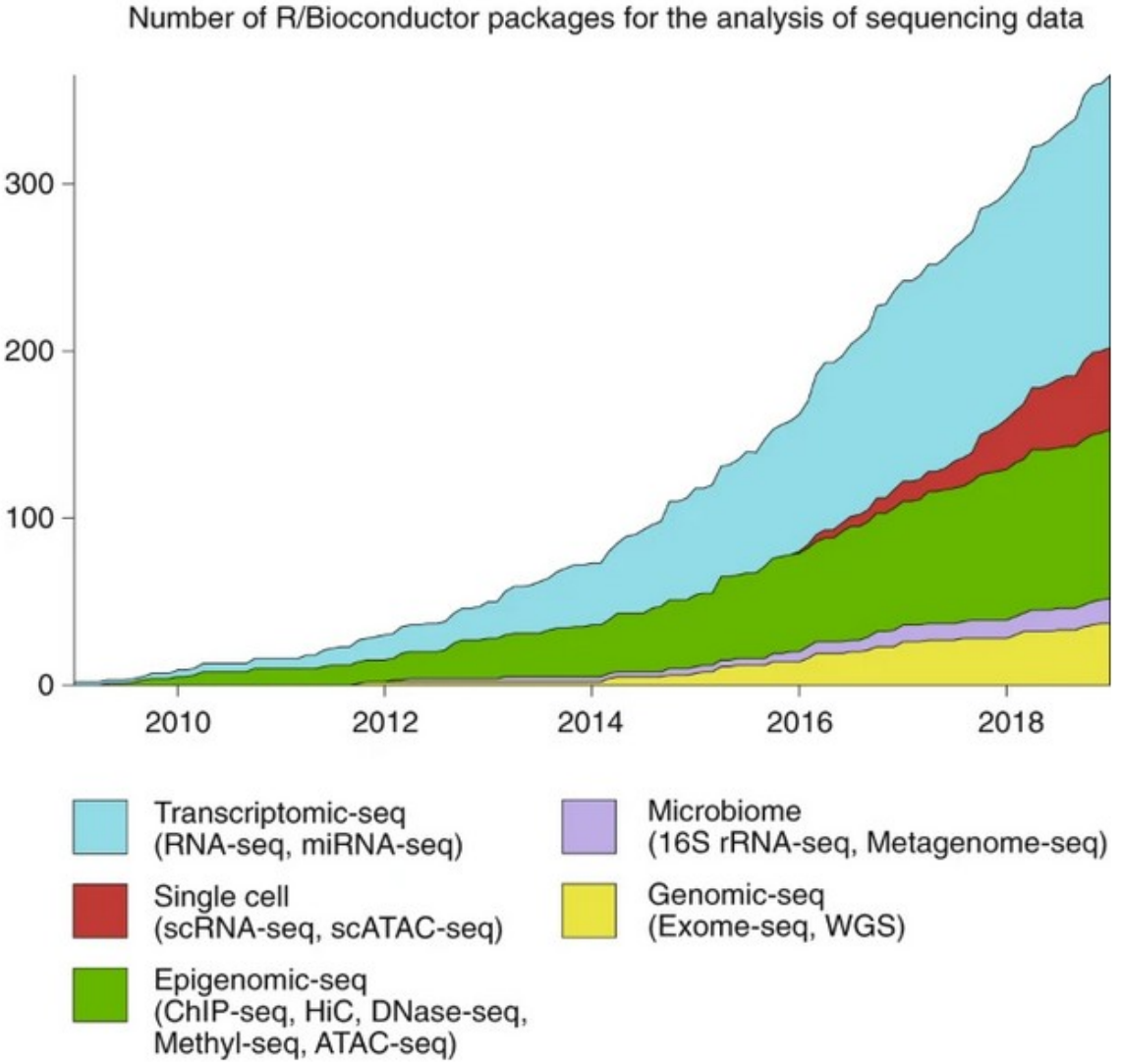


Orchestrating single-cell analysis with Bioconductor

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