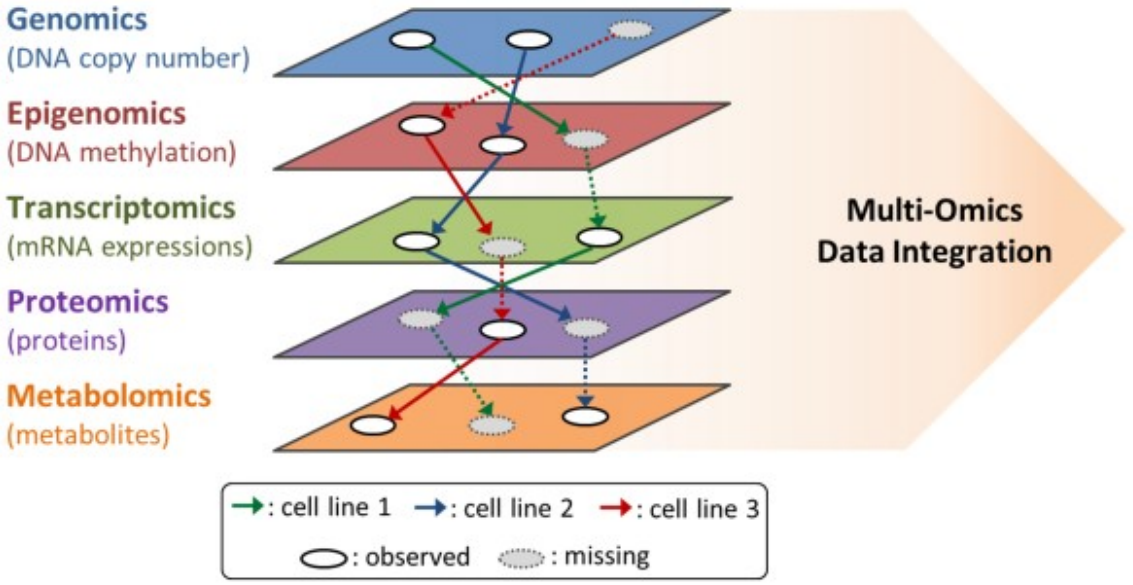
Data integration methods

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



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/

(some) topics in data integration

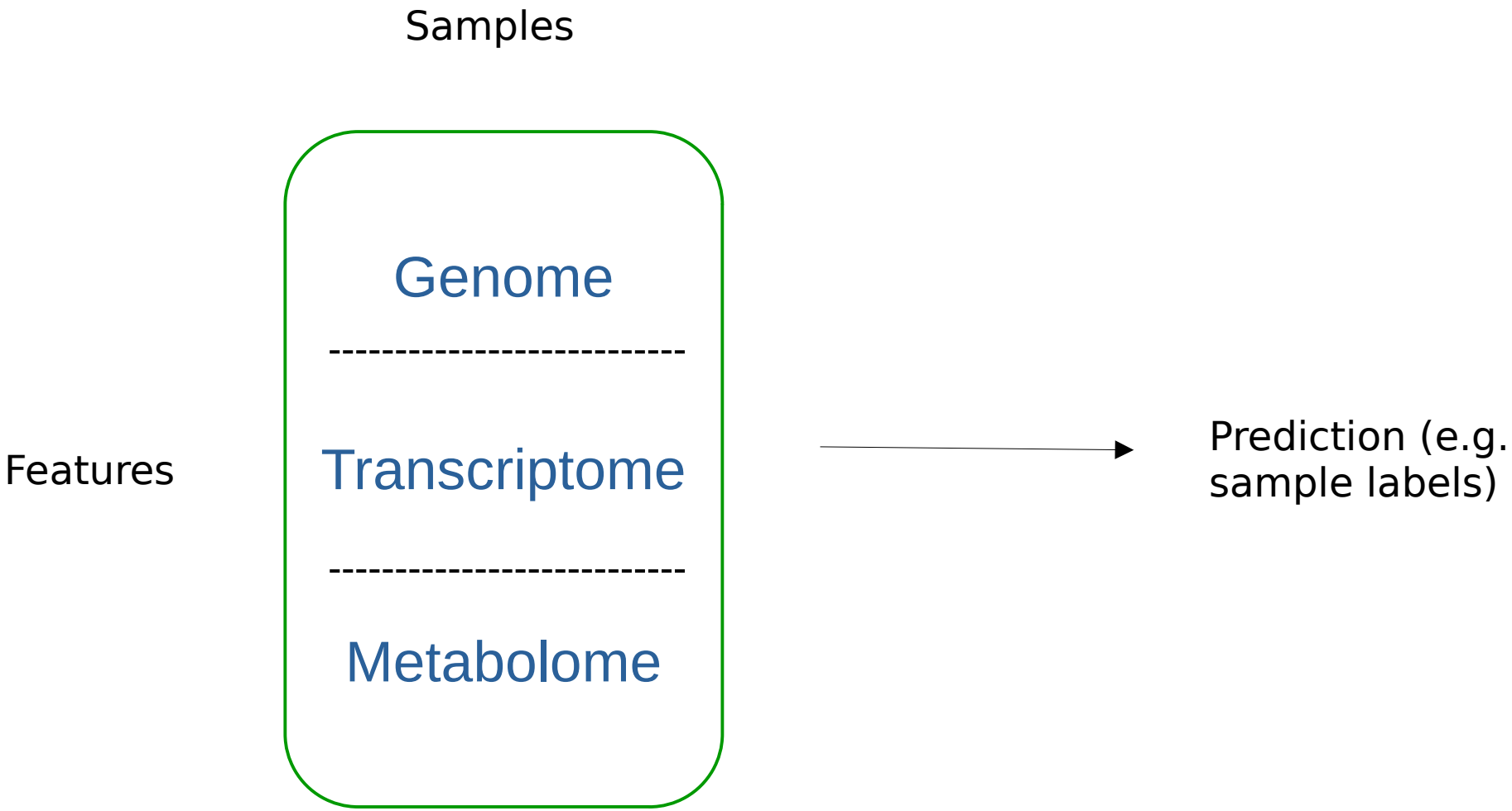
- improve predictions (of external labels)
- explore associations (btw. two or more data sets)
- identify latent processes and mechanisms
- incorporate prior knowledge

Associating data with external variables

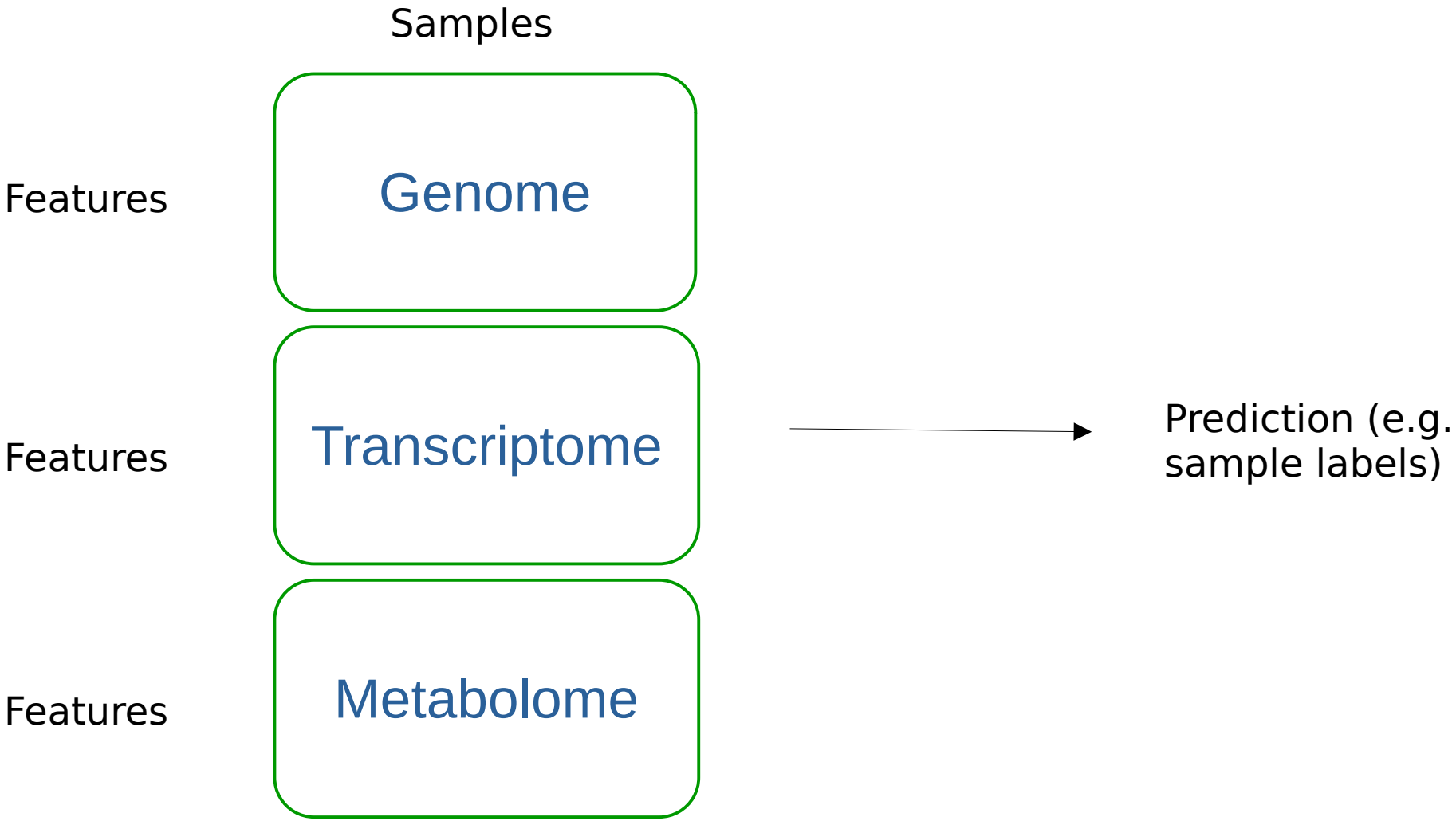
- e.g. prediction tasks

Prediction with a single data source

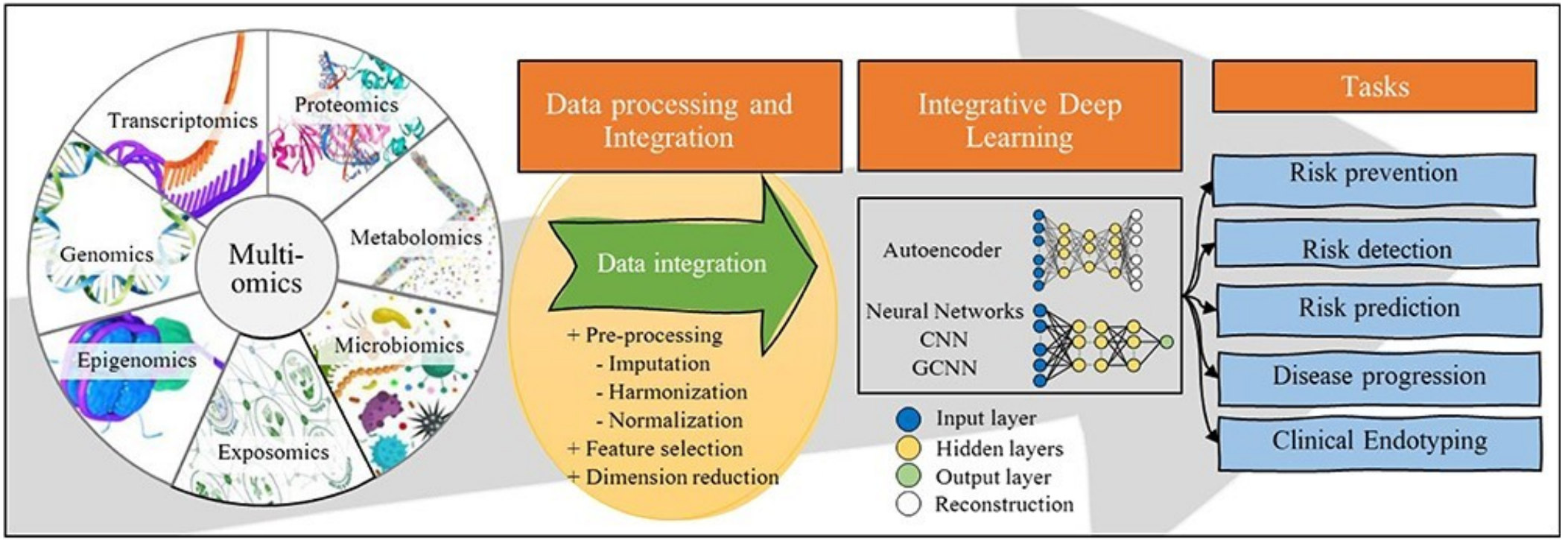
Concatenation: a null model for data integration?



Multi-view learning: advanced models for data integration?



Deep learning as an example



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]

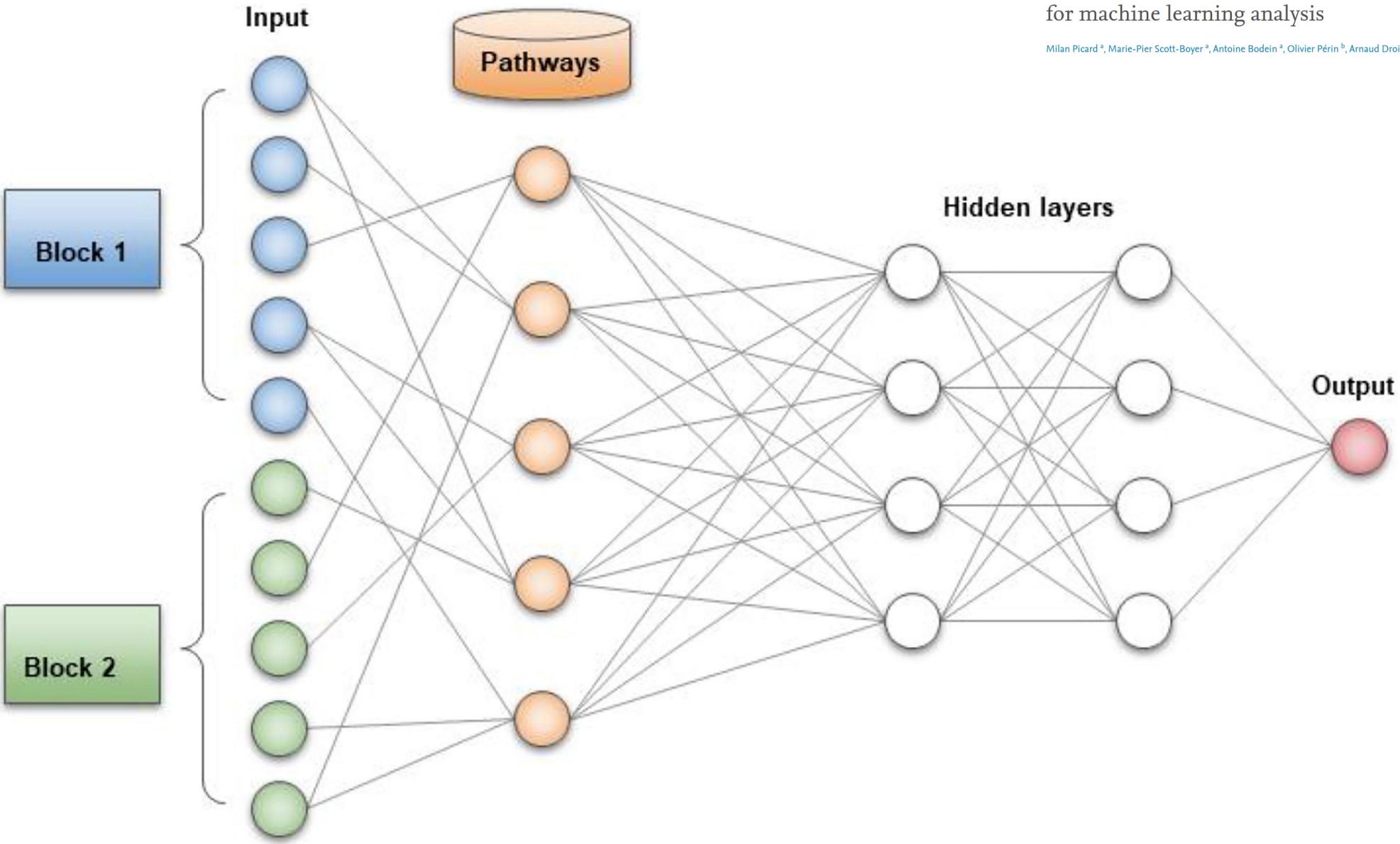
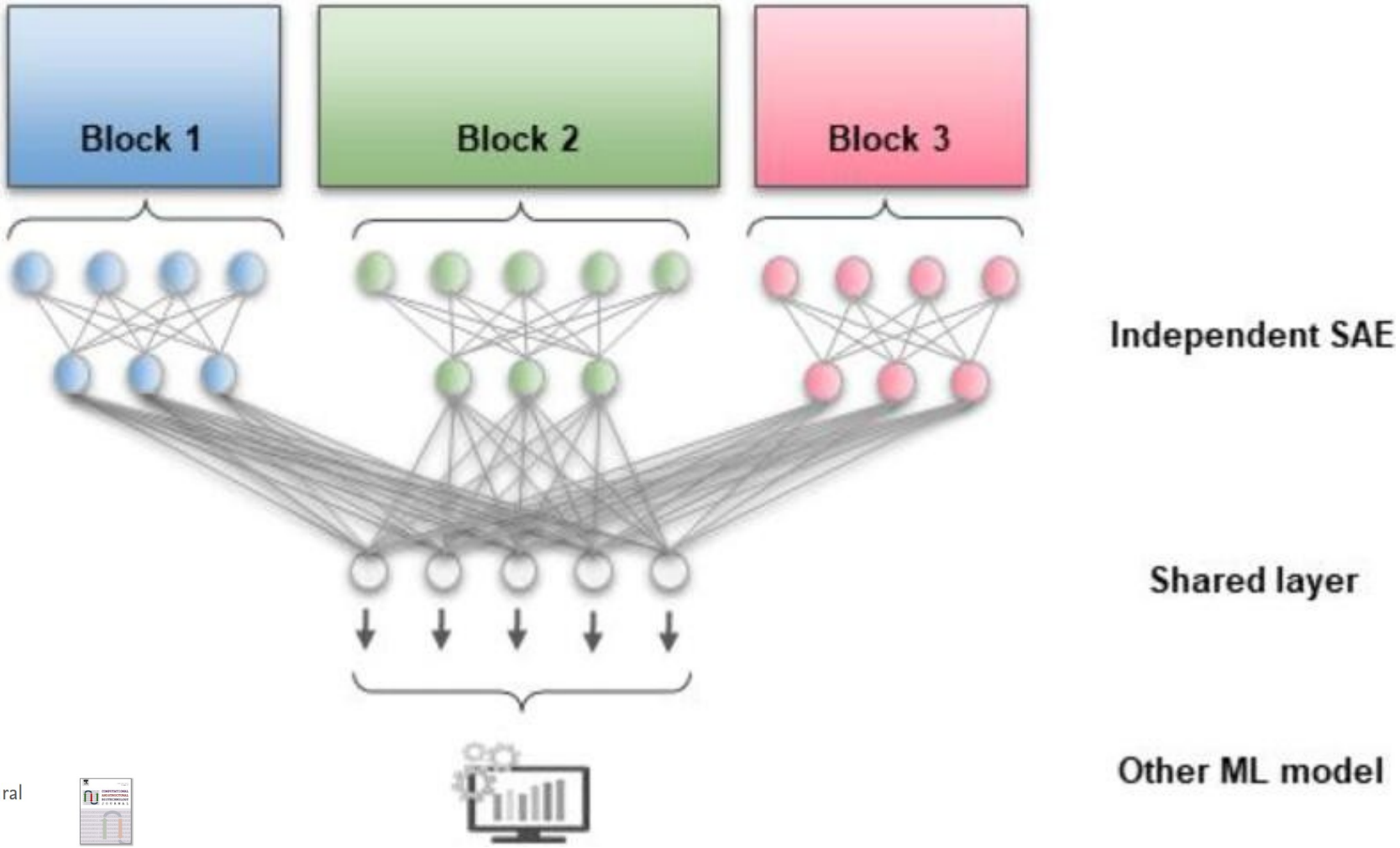


Fig. 1. Structure of an interpretable artificial neural network. The input layer is followed by an additional pathway layer, where each node corresponds to a known molecular pathway. If a molecule is known to be involved in a pathway, a connection is made between the two. Hence, important pathways implicated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

Mixed artificial 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]

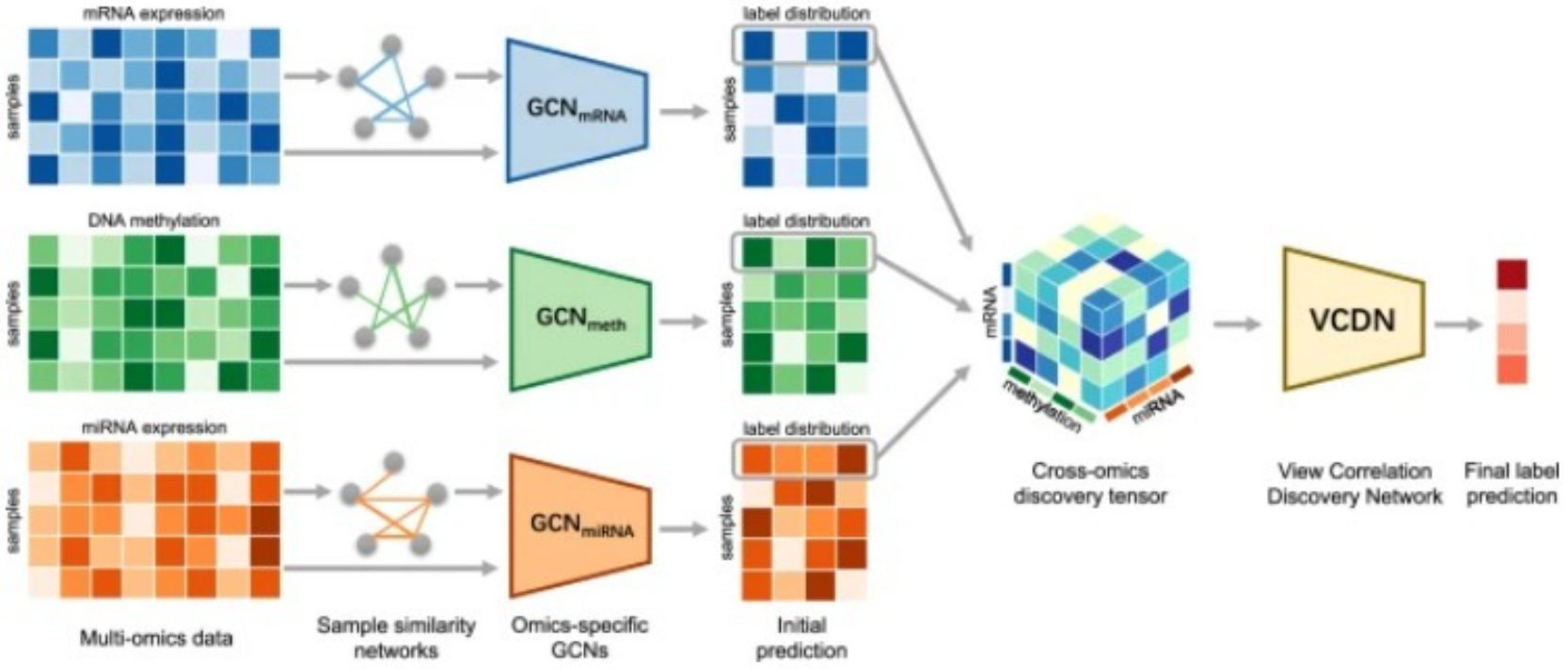
Example of a mixed artificial neural network. Each omics block is first reduced to a latent representation using independent Stacked Sparse Autoencoders (SAE). The new representations learned are integrated in a final shared layer. The common representation is used for downstream analysis such as prediction or clustering. Figure inspired from Xu et al. (2019)

Improving predictions by data integration

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

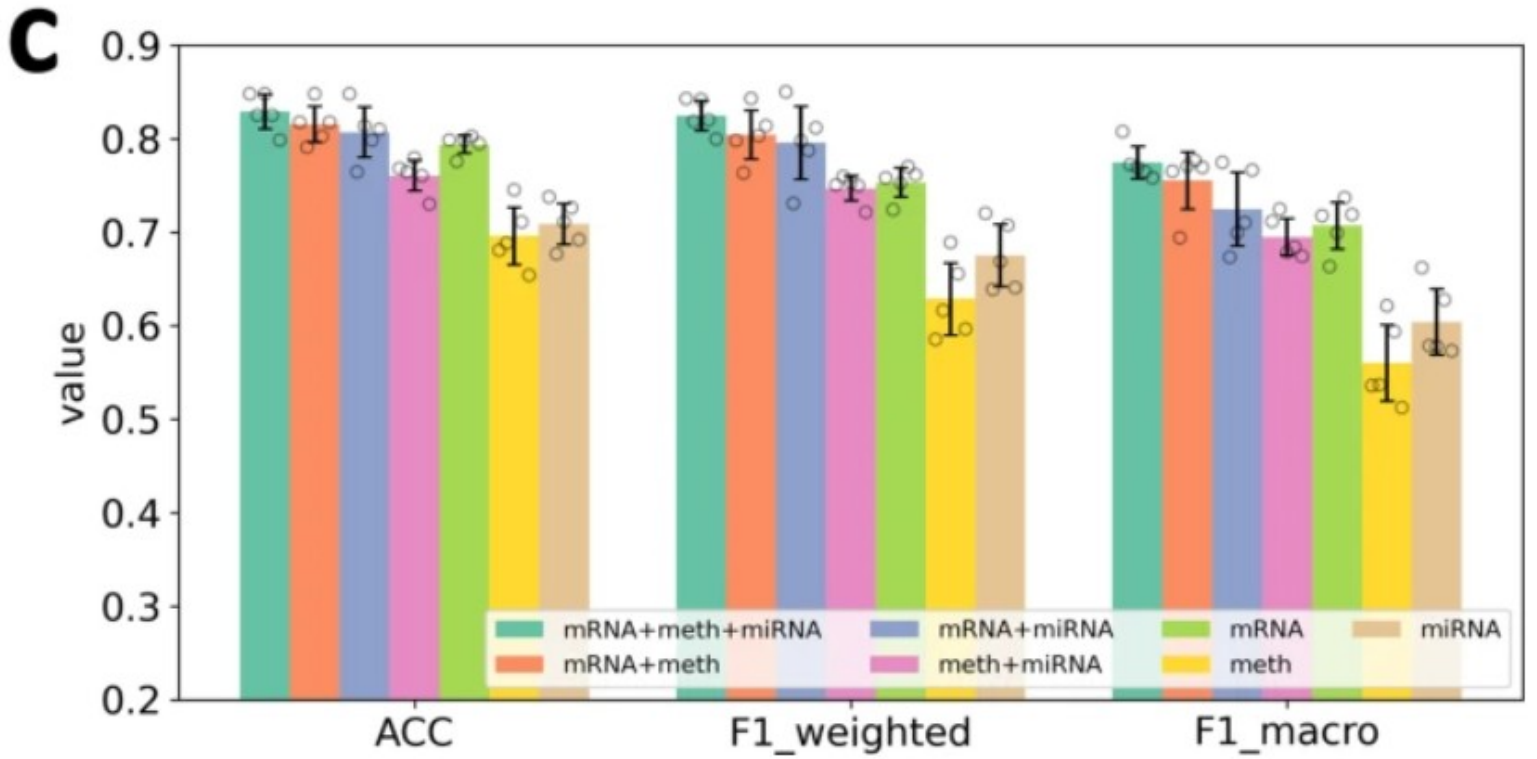
Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding & Kun Huang

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

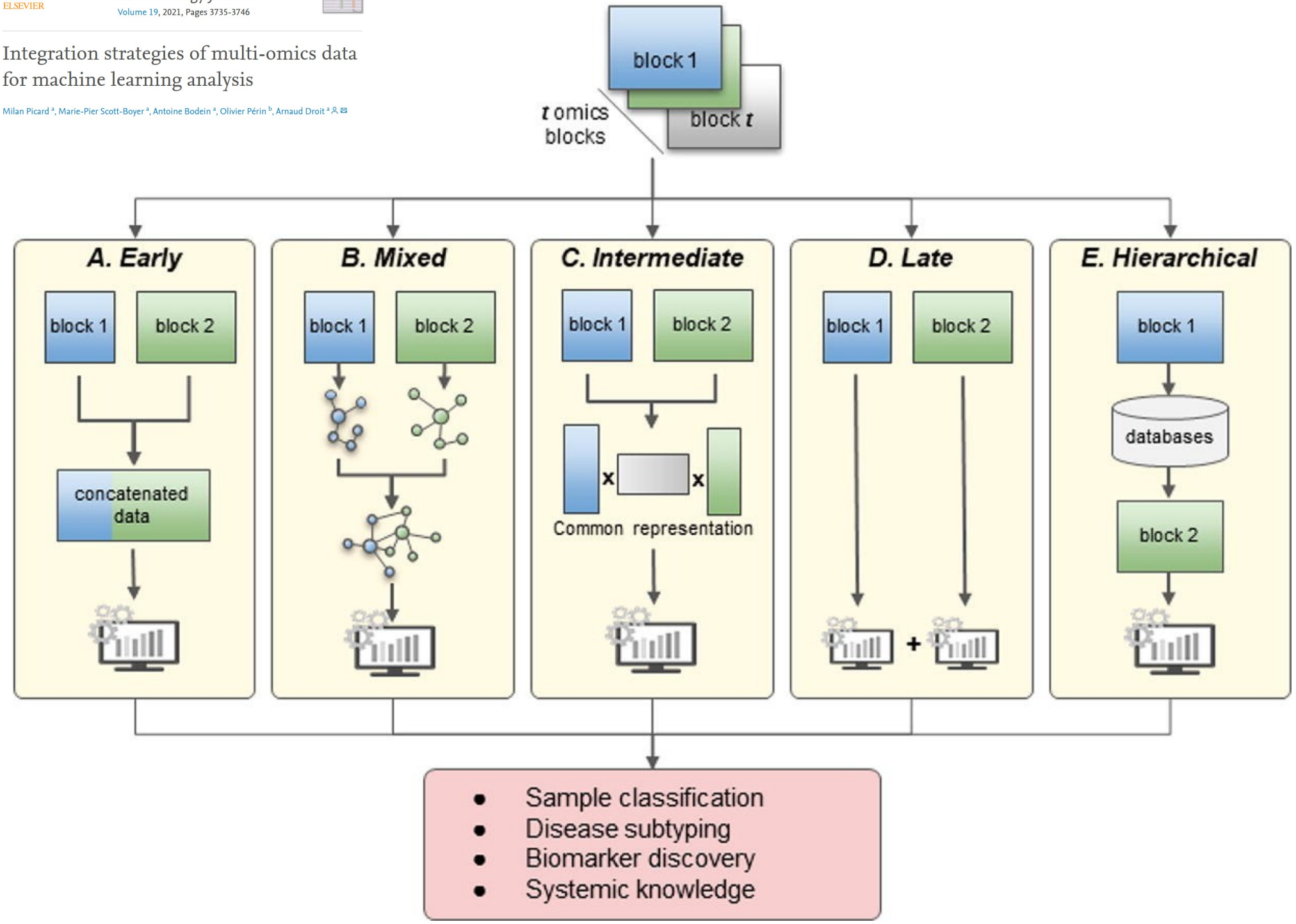
Integration strategies of multi-omics data for machine learning analysis

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?

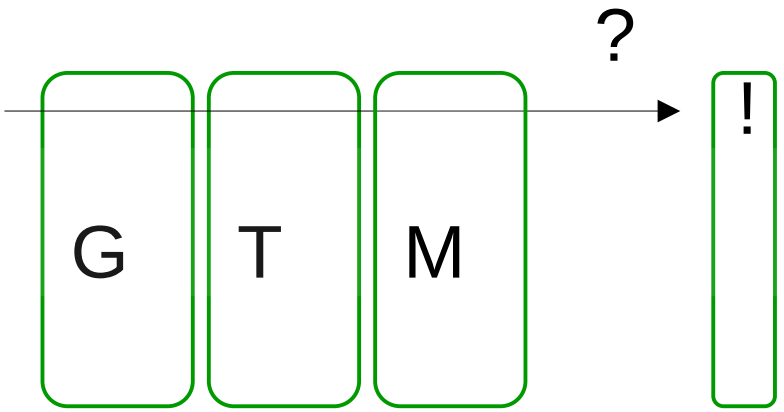
- **scale**: small mechanistic models vs. large-scale exploration?

- **two or more data sets**?

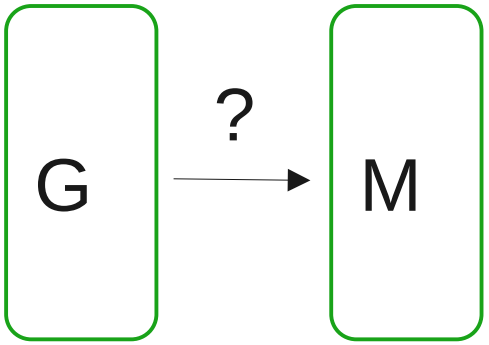
- **known structure**?

- **computational requirements**?

Predicting external labels

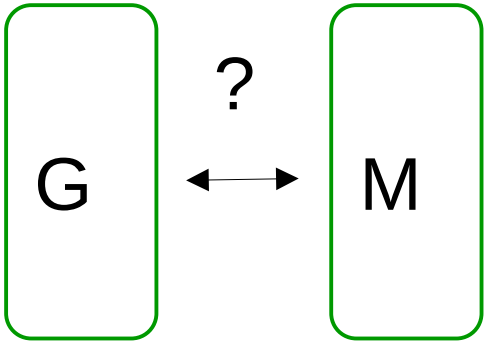


Association -
one data set is primary

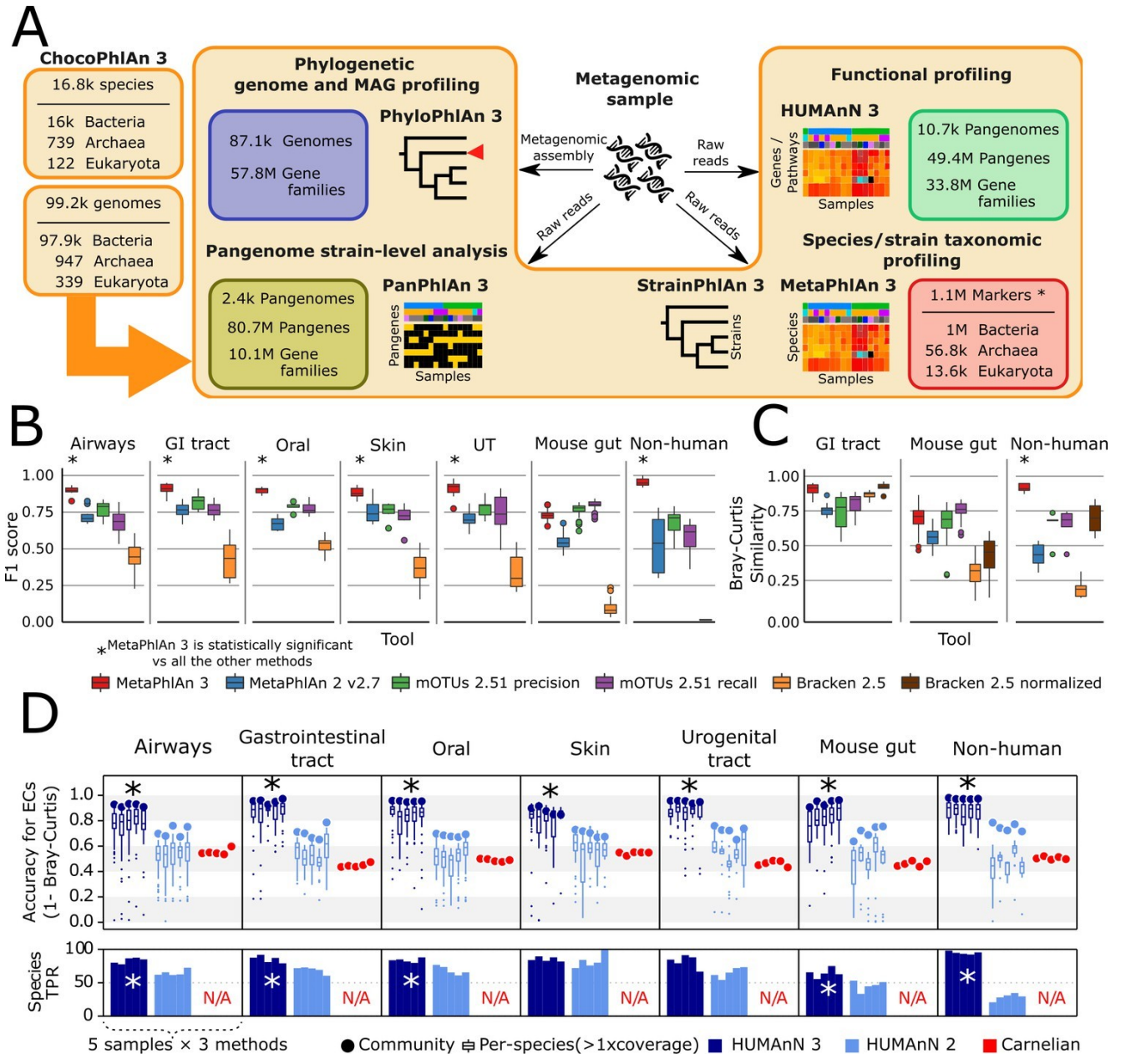


Association -

all data sets are equal



HUMAnN is a method for profiling the abundance of microbial metabolic pathways from metagenomic or metatranscriptomic sequencing data

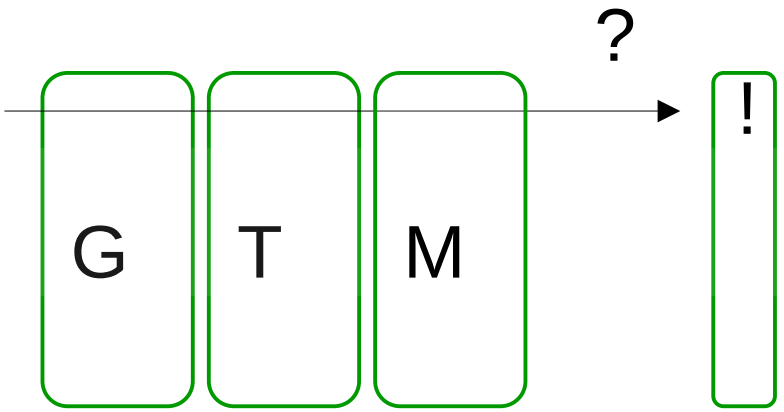


Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3

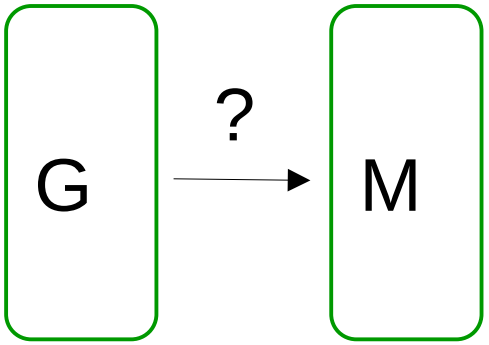
Francesco Beghini, Lauren J McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, Paolo Manghi, Matthias Scholz see all »

Department CIBIO, University of Trento, Italy; Harvard T.H. Chan School of Public Health, United States; The Broad Institute of MIT and Harvard, United States; Department of Food Quality and Nutrition, Research and Innovation Center, Edmund Mach Foundation, Italy; IEO, European Institute of Oncology IRCCS, Italy

Predicting external labels

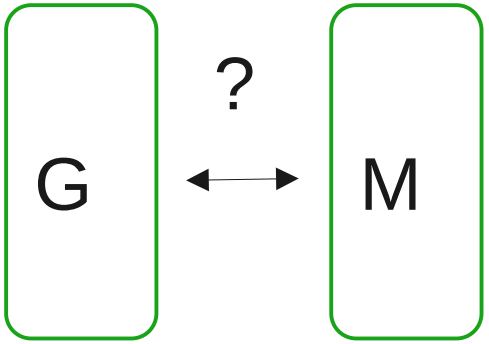


Association -
one data set is primary



Association -

all data sets are equal



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

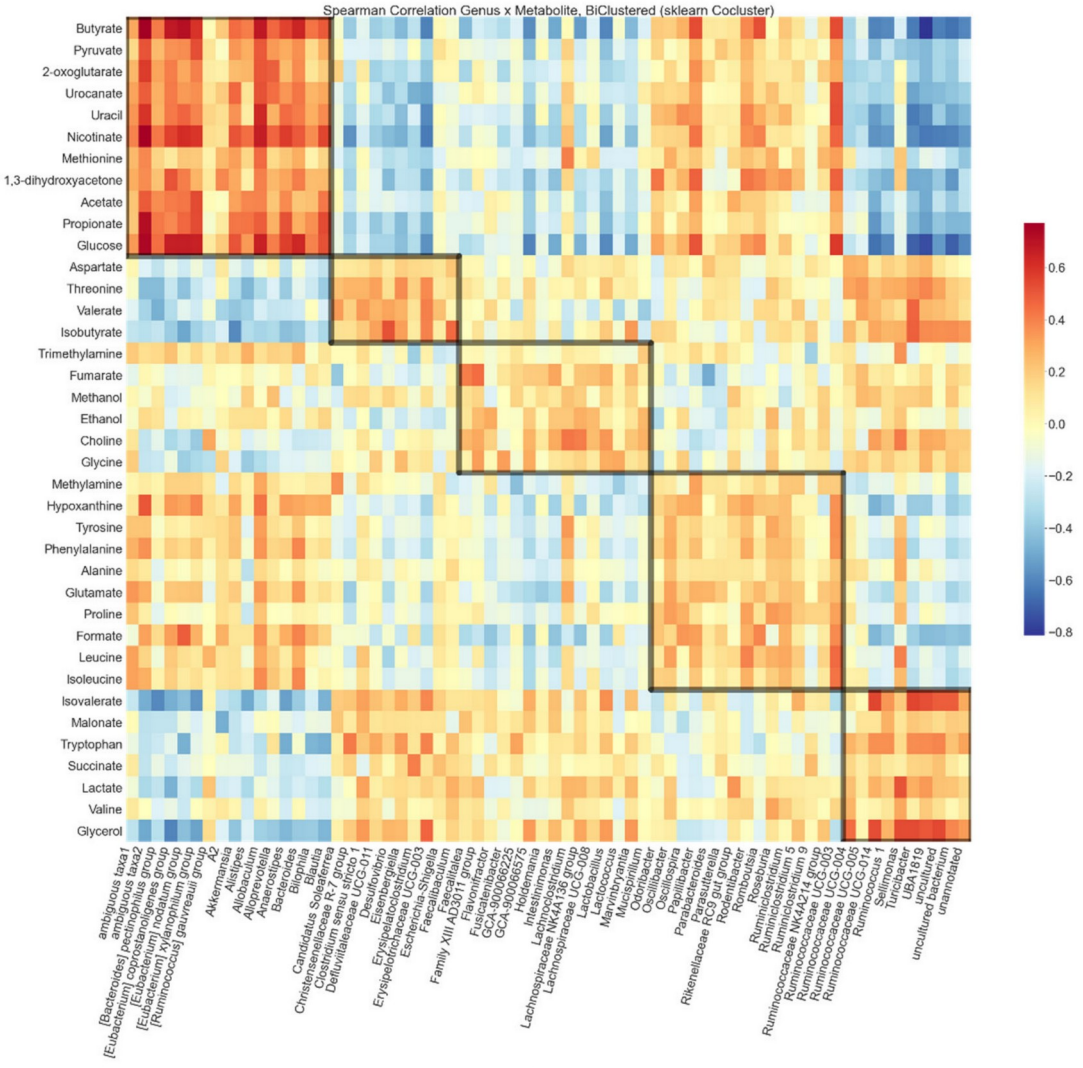
Open Access Article

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

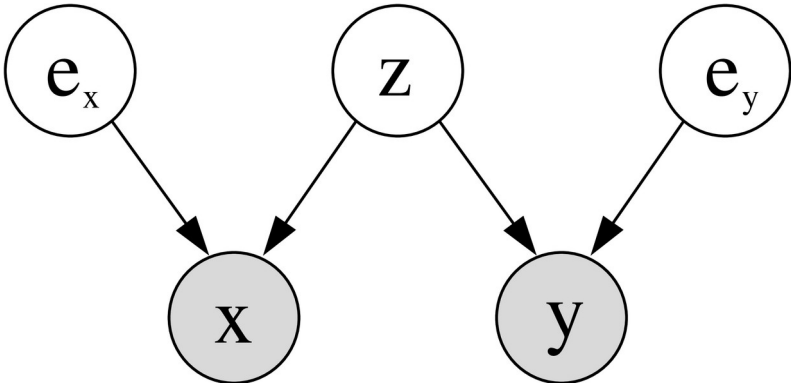
How to generalize to more than two data sets..?



(a)

Multi-view learning

- Symmetric setup

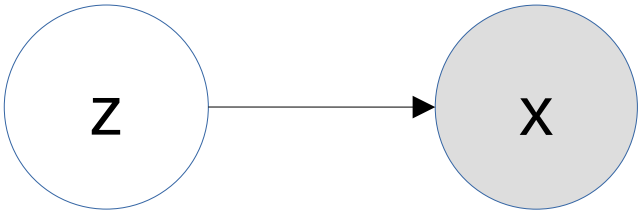


Example: PCA vs. CCA

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



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

Human gut microbiome ordination

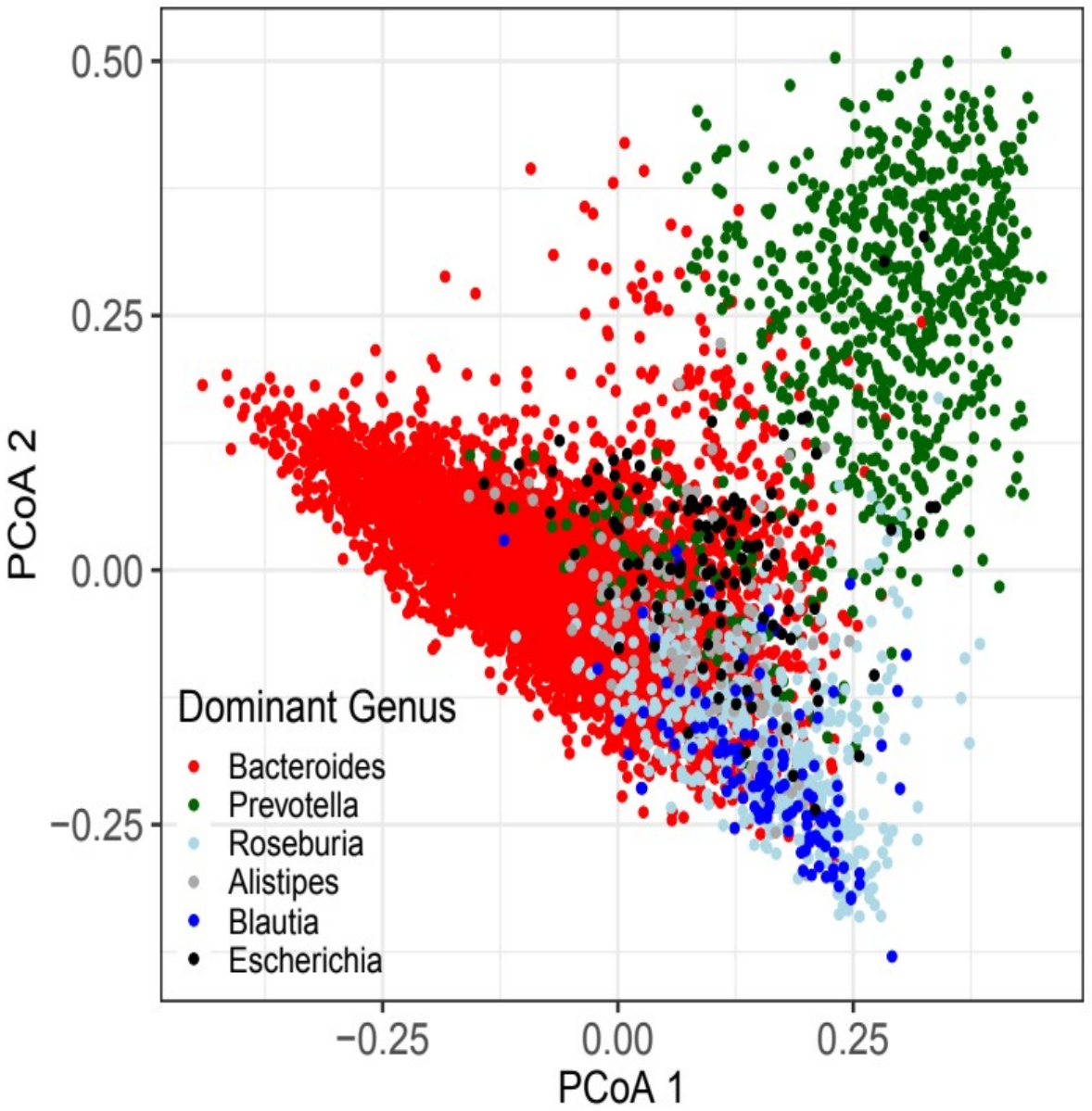
$$X = W_x \mathbf{z} + \varepsilon_x$$

Z: cluster of each individual (“color”)

ε_x: Noise

W: Cluster profile

X: Observed point



Article | Open Access | Published: 11 May 2021

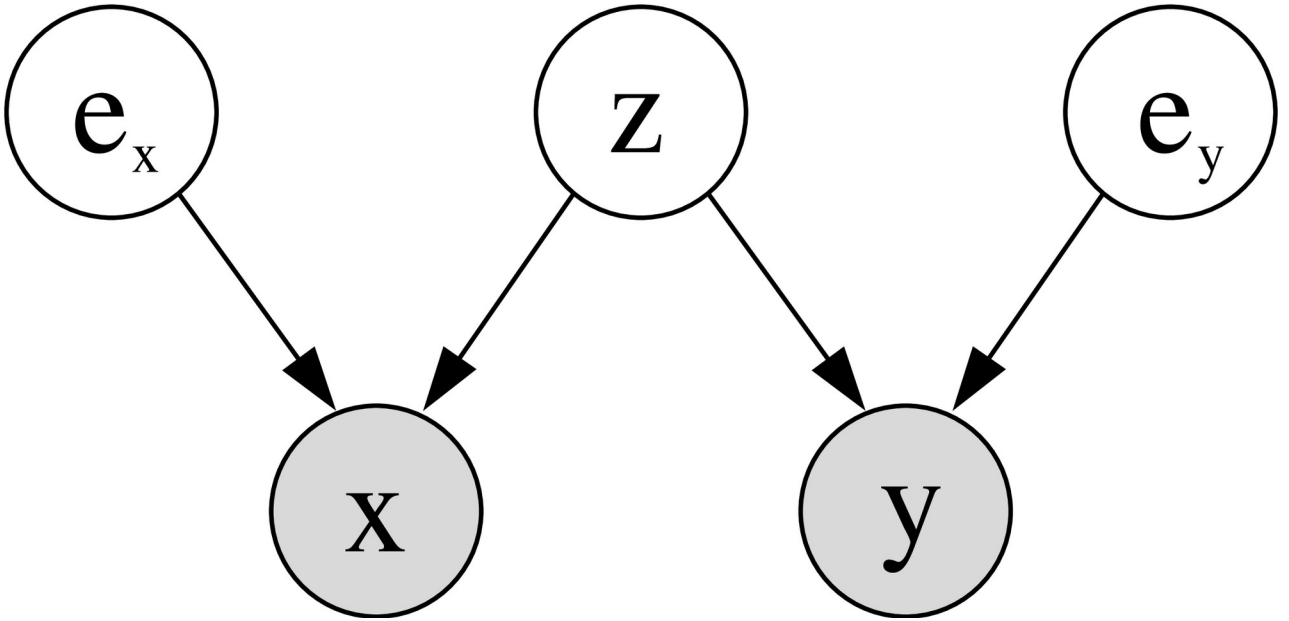
Taxonomic signatures of cause-specific mortality risk in human gut microbiome

Aaro Salosensaari, Ville Laitinen, Aki S. Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, Georg Alfthan, Michael Inouye, Jeramie D. Watrous, Tao Long, Rodolfo A. Salido, Karenina Sanders, Caitriona Brennan, Gregory C. Humphrey, Jon G. Sanders, Mohit Jain, Pekka Jousilahti, Veikko Salomaa, Rob Knight, Leo Lahti & Teemu Niiranen

Nature Communications 12, Article number: 2671 (2021) | Cite this article

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

Multi-view learning:
CCA is a generalization of PCA
(Bach & Jordan 2005)



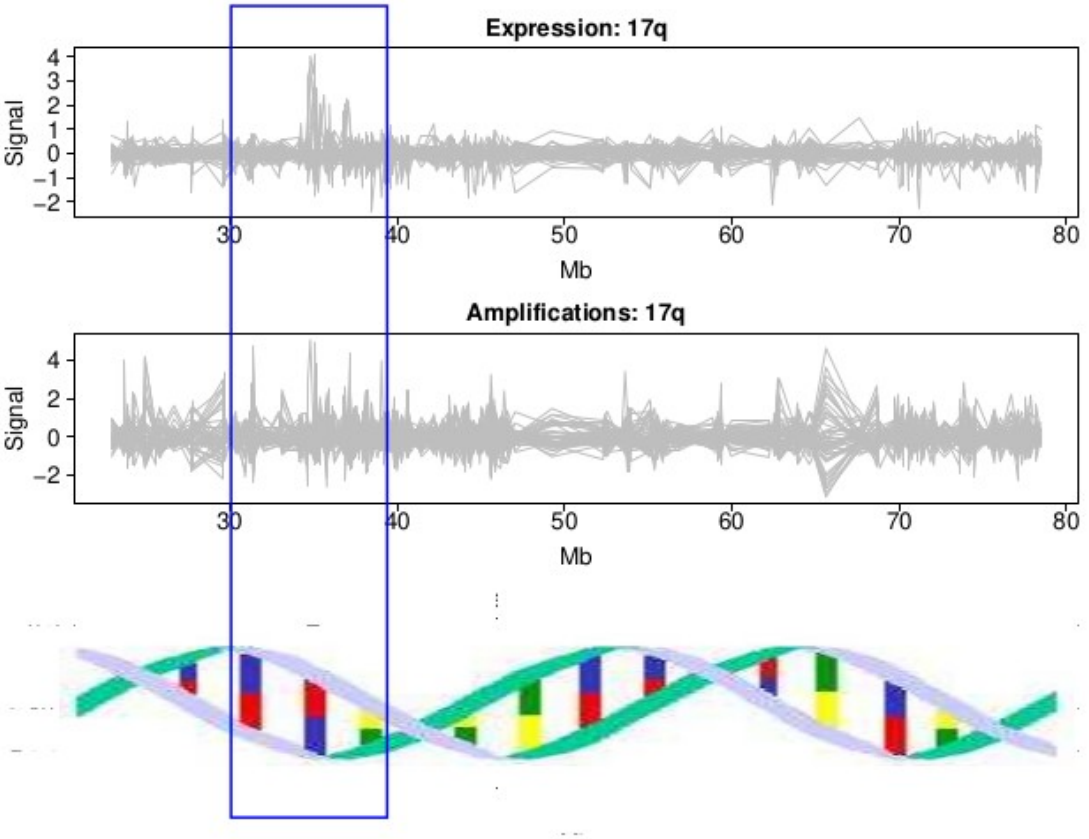
https://github.com/mblstamps/stamps2019/blob/master/STAMPS2019_overview_Pop.pdf

Chromosome arm 17q

Investigate dependencies within local chromosomal regions using sliding window

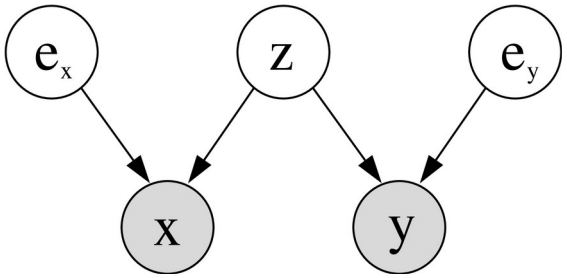
X

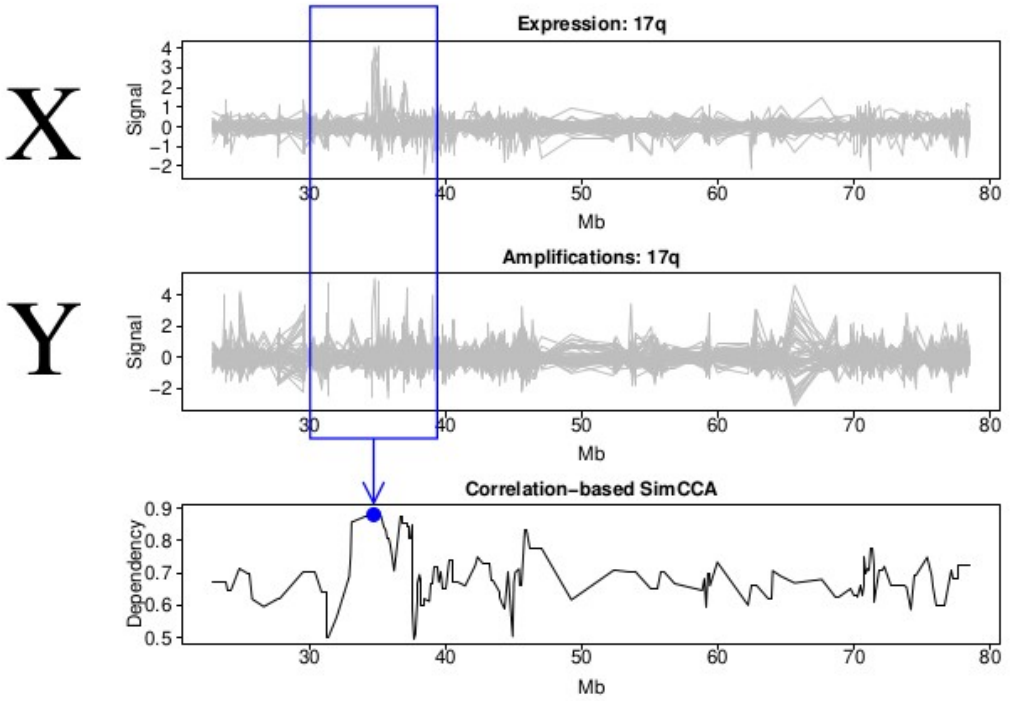
Y



Chromosome arm 17q: results

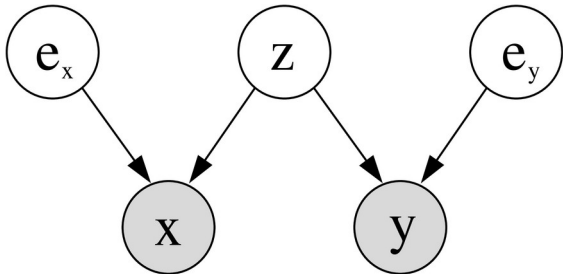
SimCCA measures dependency between data sources within each chromosomal region

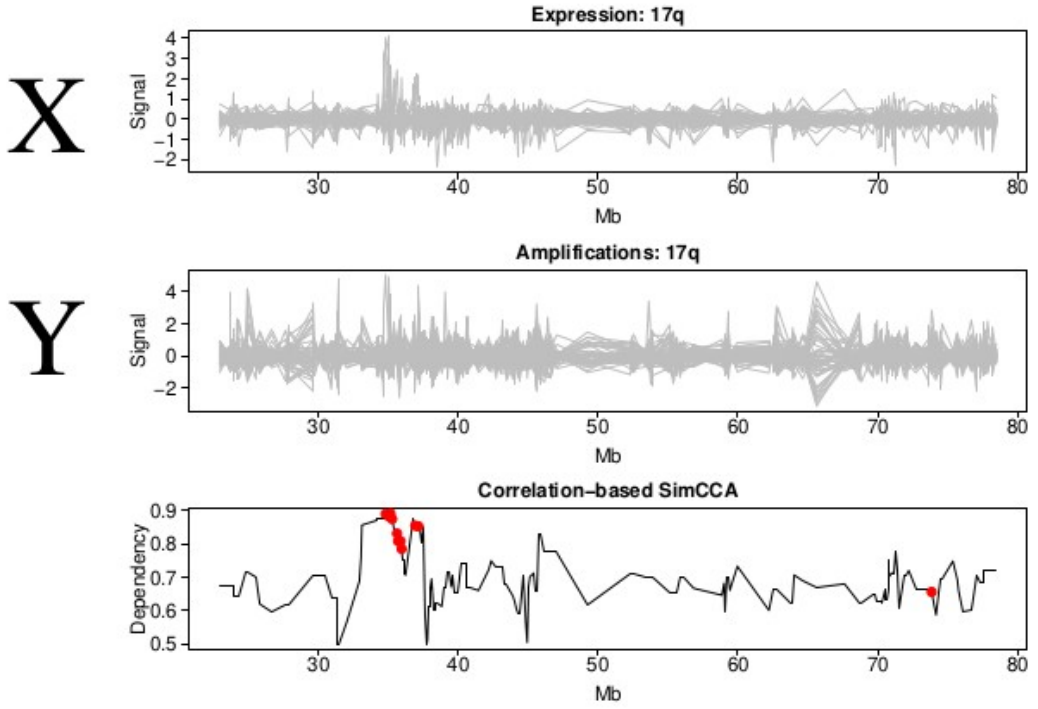




Chromosome arm 17q: results

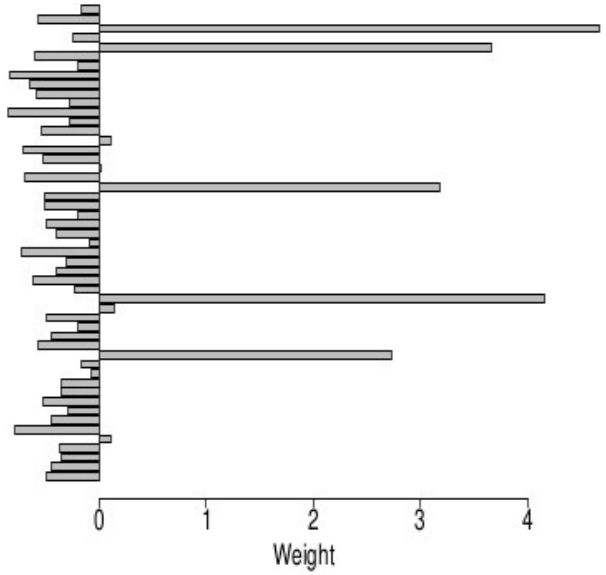
SimCCA reveals known gastric cancer-associated chromosomal regions



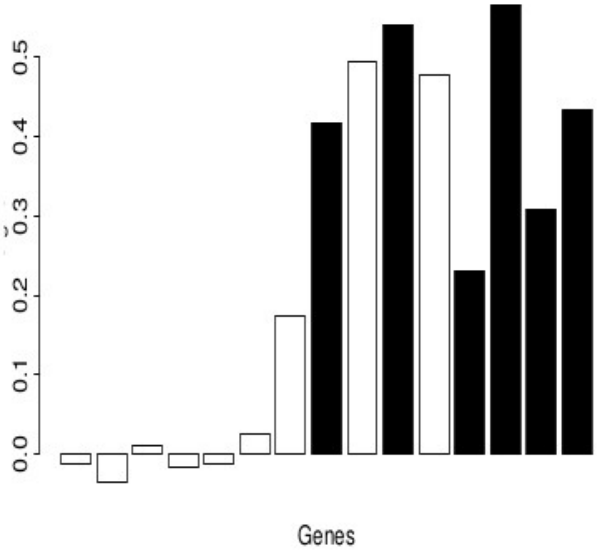


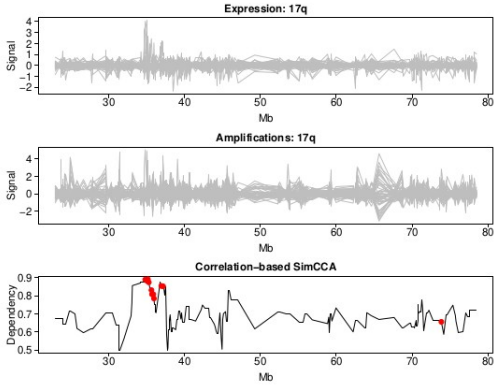
Interpreting the parameters

Z: affected patients



W: dependent observations





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

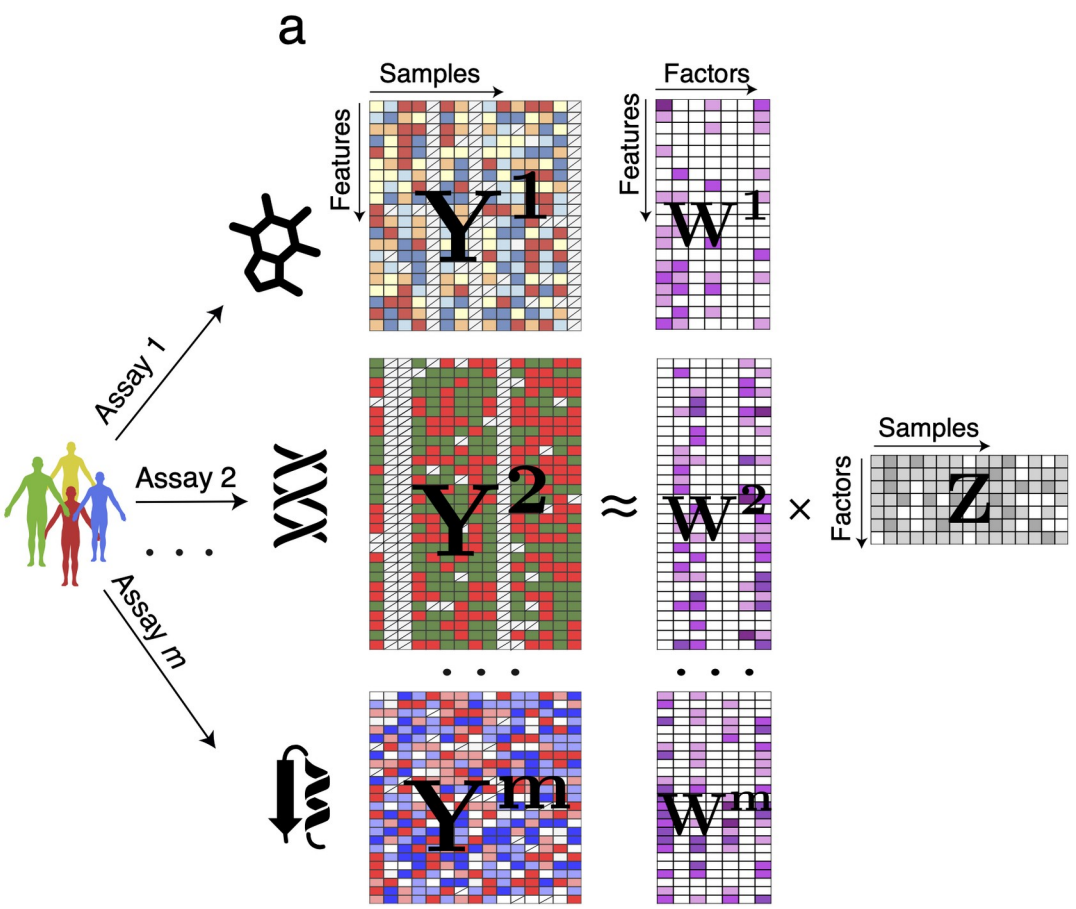
Method | 20 June 2018 | OPEN ACCESS TRANSPARENT PROCESS

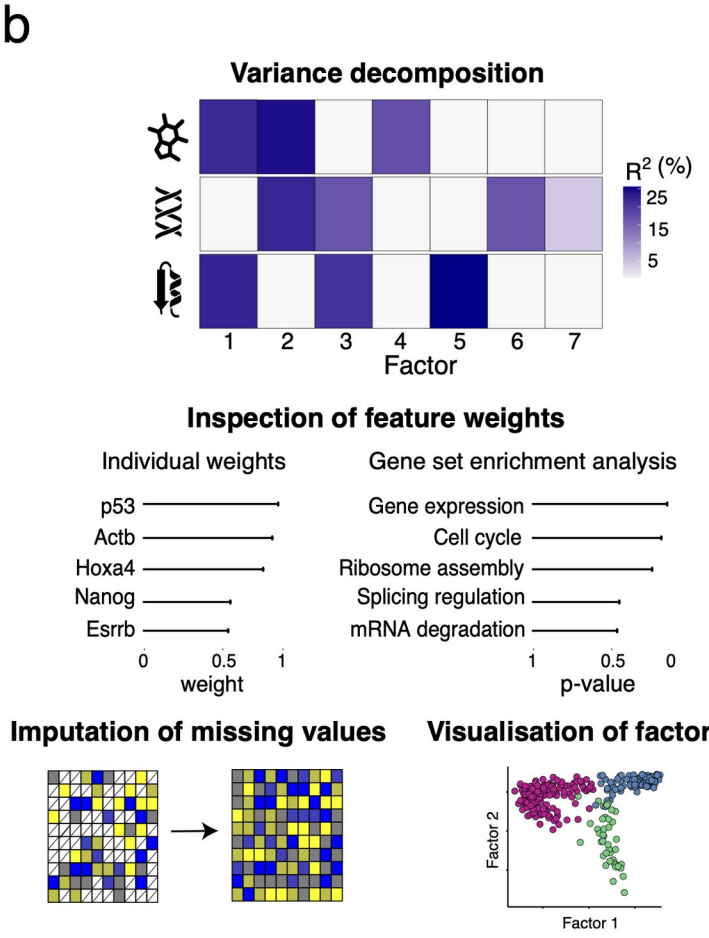
Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets

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

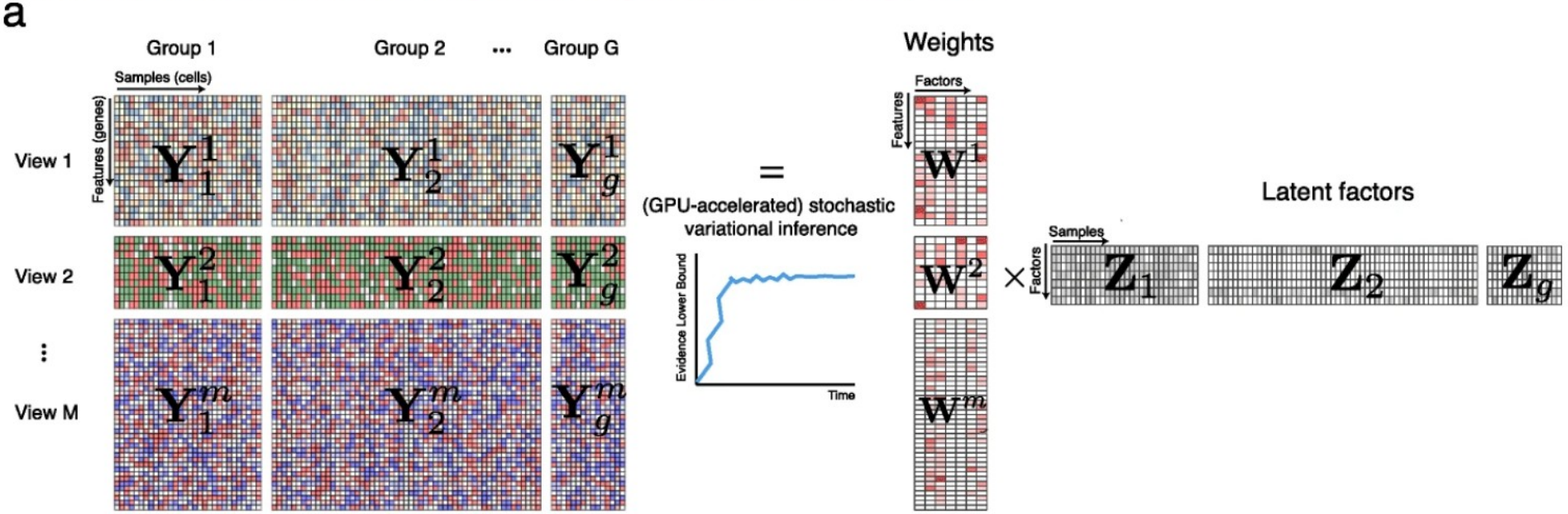
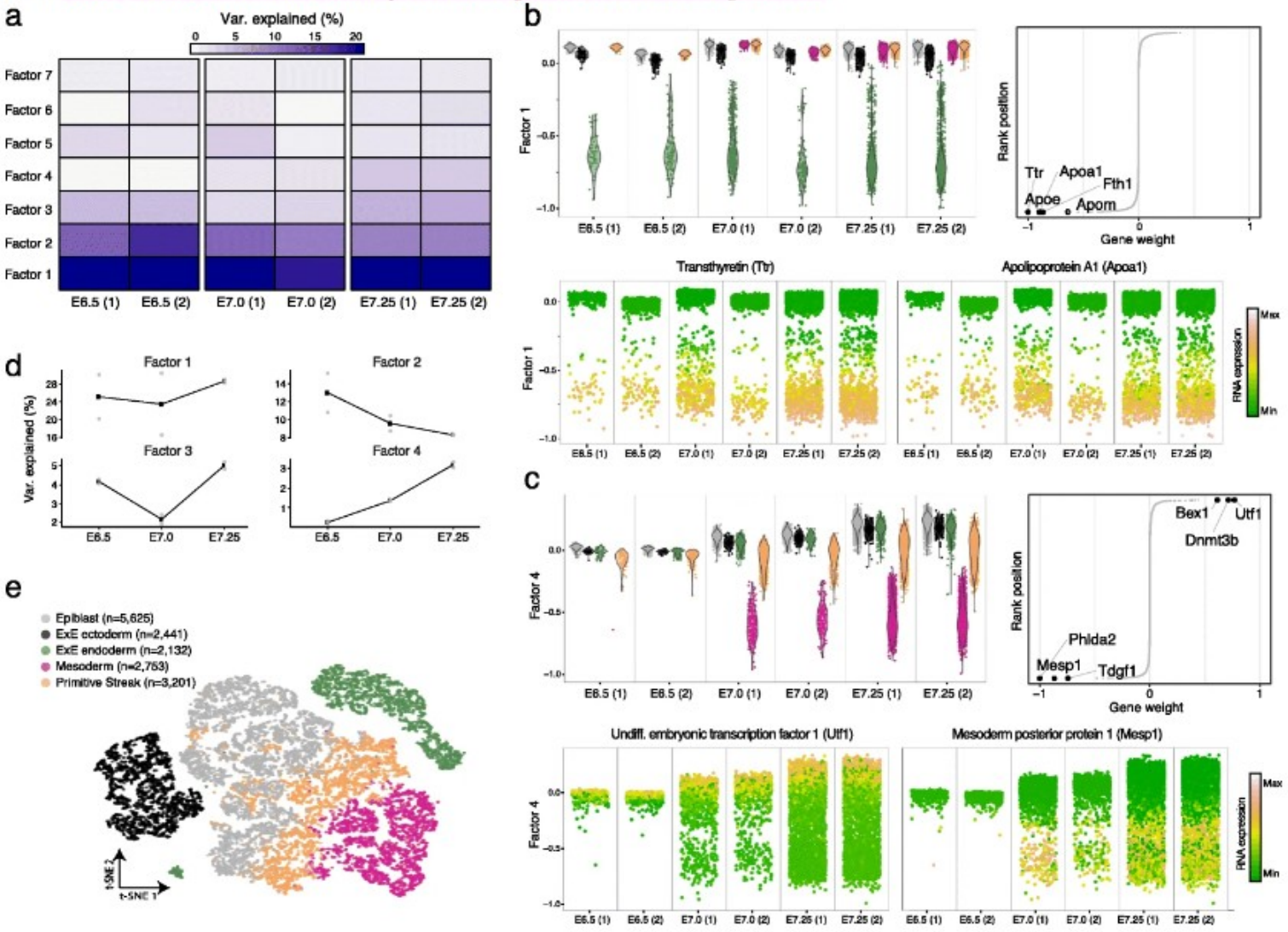


Fig. 2

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

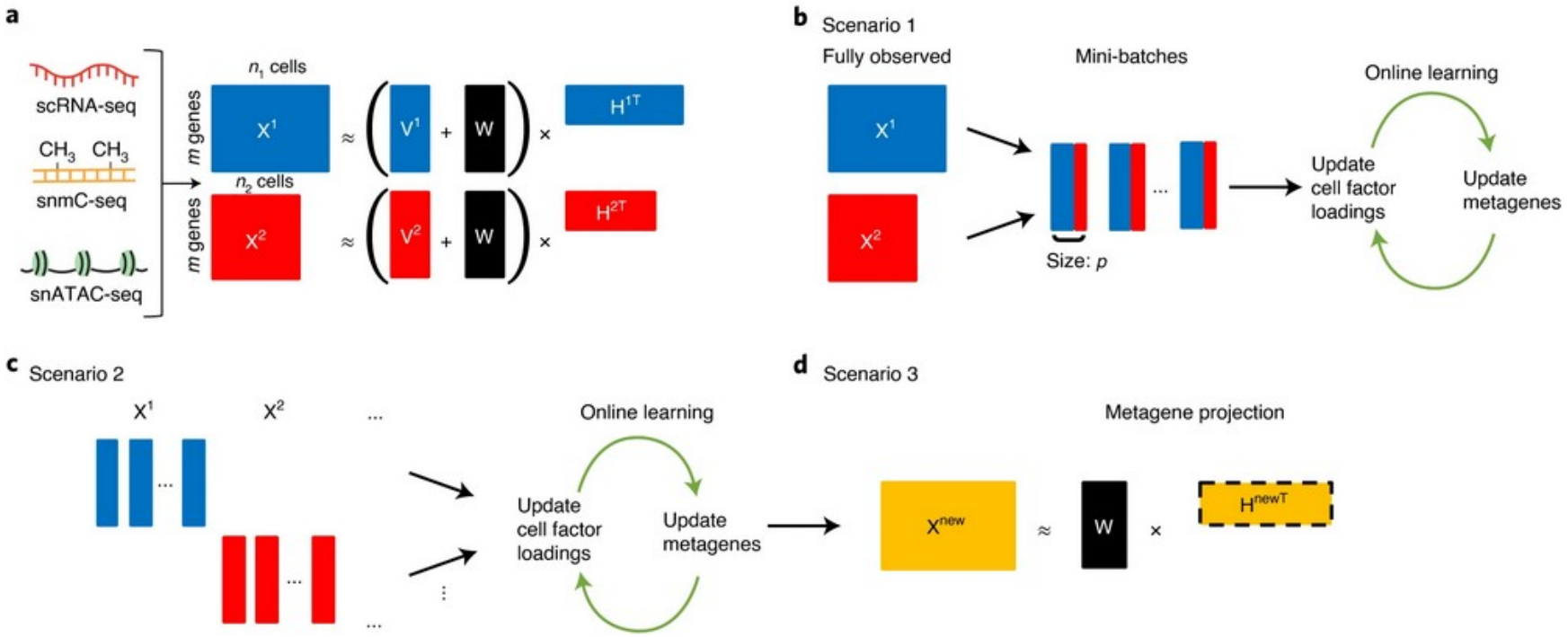
Chao Gao, Jialin Liu, April R. Kriebel, Sebastian Preissl, Chongyuan Luo, Rosa Castanon, Justin Sandoval, Angeline Rivkin, Joseph R. Nery, Margarita M. Behrens, Joseph R. Ecker, Bing Ren & Joshua D. Welch

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.

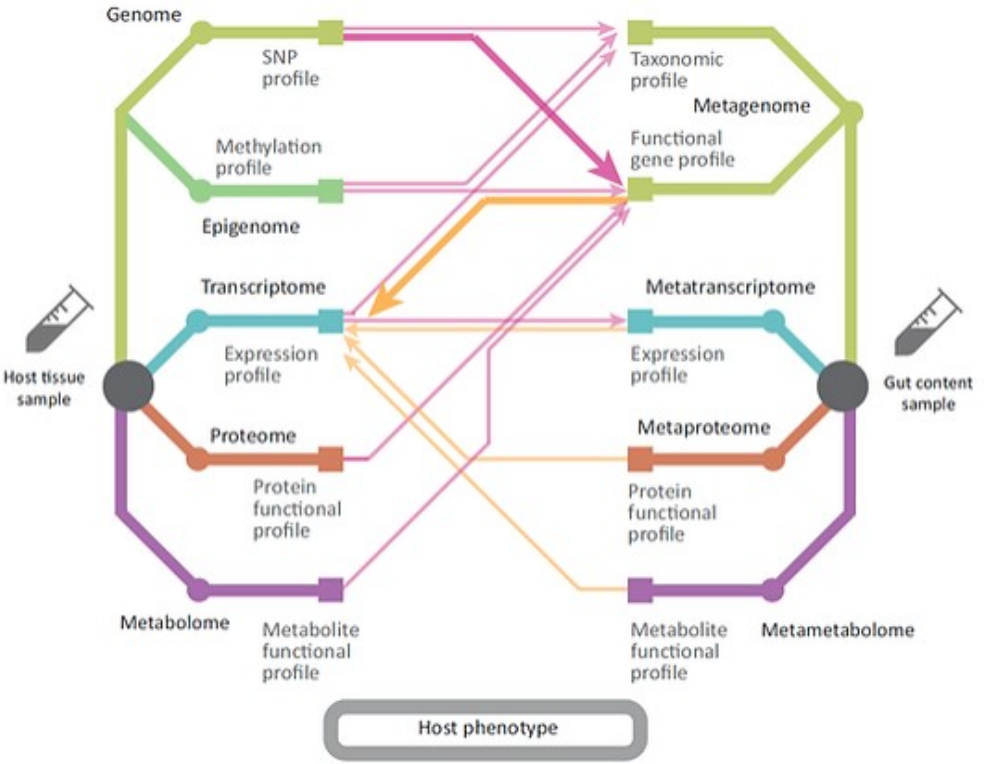
REVIEW ARTICLE
Front. Genet., 28 August 2019 | https://doi.org/10.3389/fgene.2019.00627

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.

TreeSummarizedExperiment data container

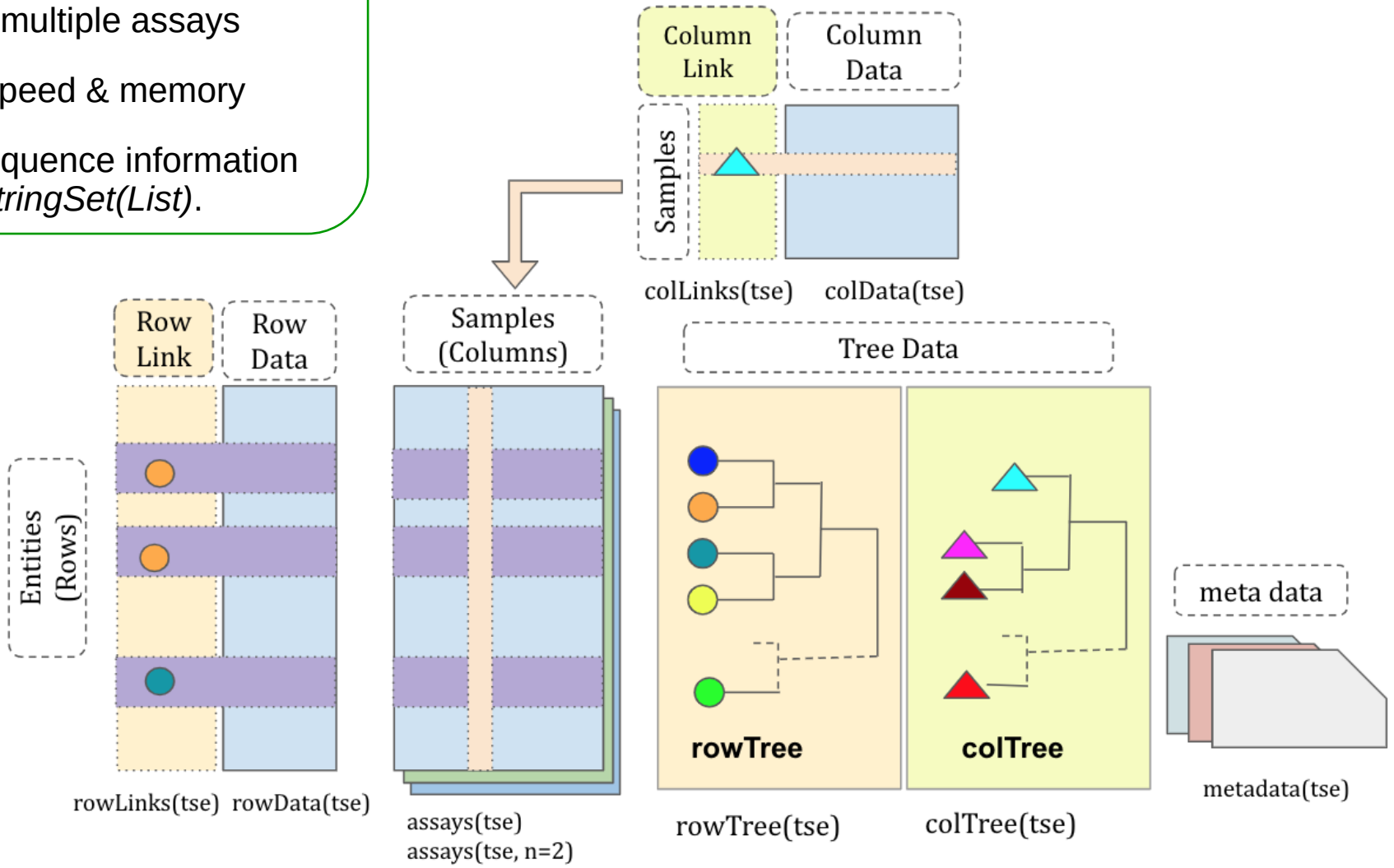
by Ruizhu @fiona Huang; initially proposed for microbiome research by Hector Bravo & Domenick Braccia

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

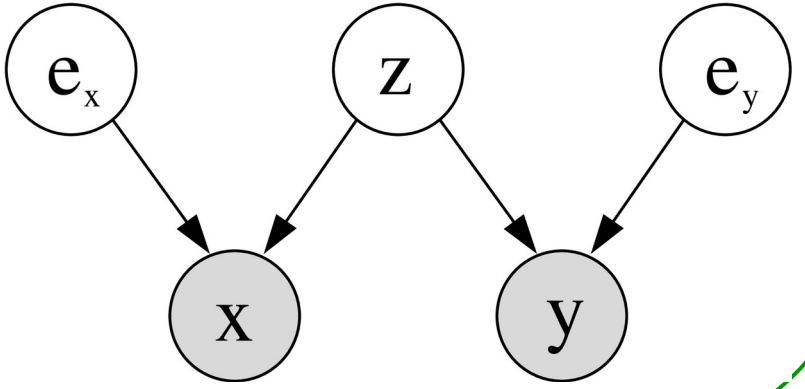
Figure 5: The online version provides the text in HTML, data files and up-to-date code.

Take-home messages

- Heterogeneity of problems

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

- Importance of study question



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)

Association analyses with biclustering

Demo on MOFA

Supervised learning: Regression and classification with random forests

Validation and interpretation of black box models