Integrating Biological Data (Omics) via Machine Learning

CZI Online Workshop

Nikolay Oskolkov, NBIS SciLifeLab

Lund, 01.04.2022

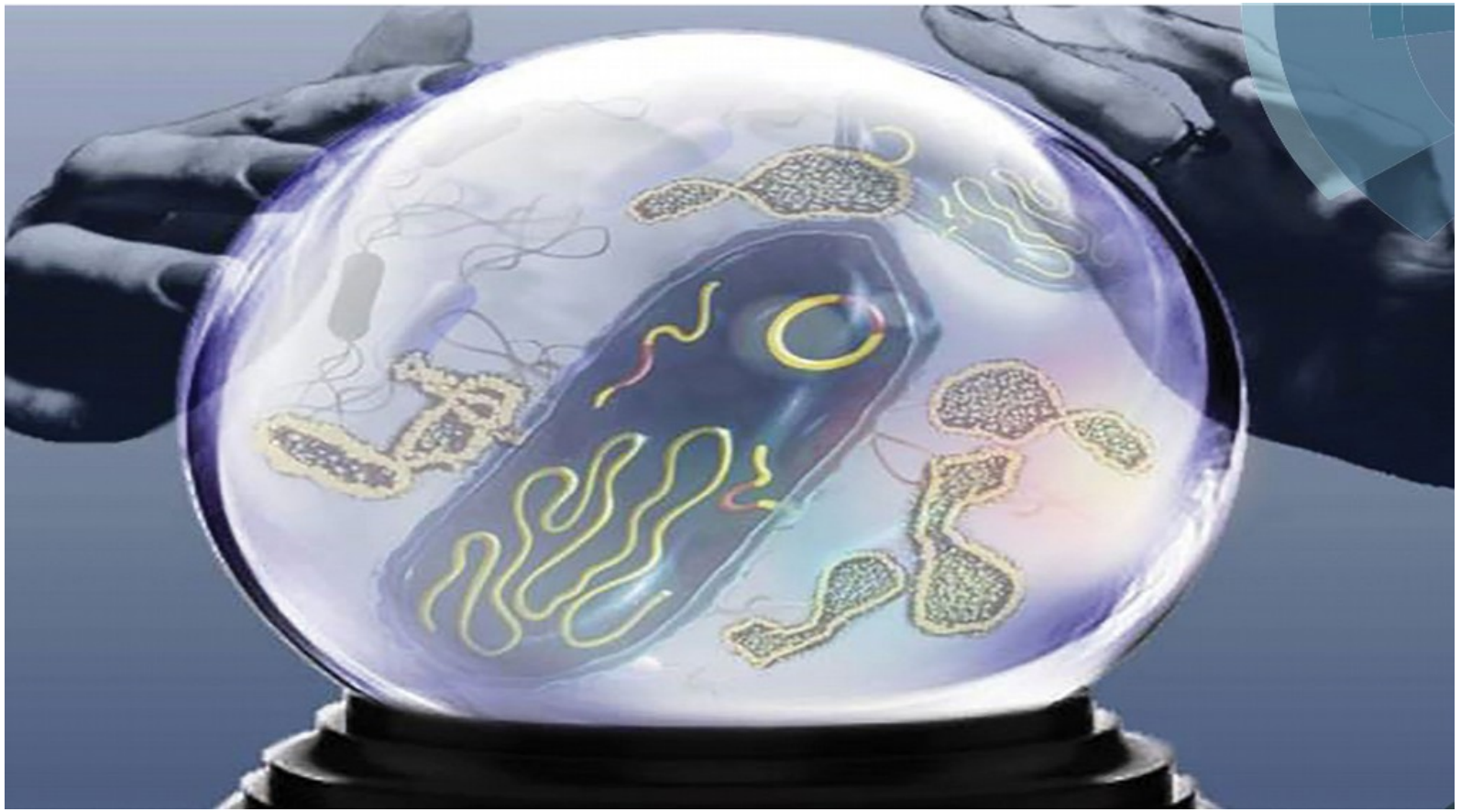


Image adapted from Molecular Omics, Issue 1, 2018

Multi-Omics Data in Biology

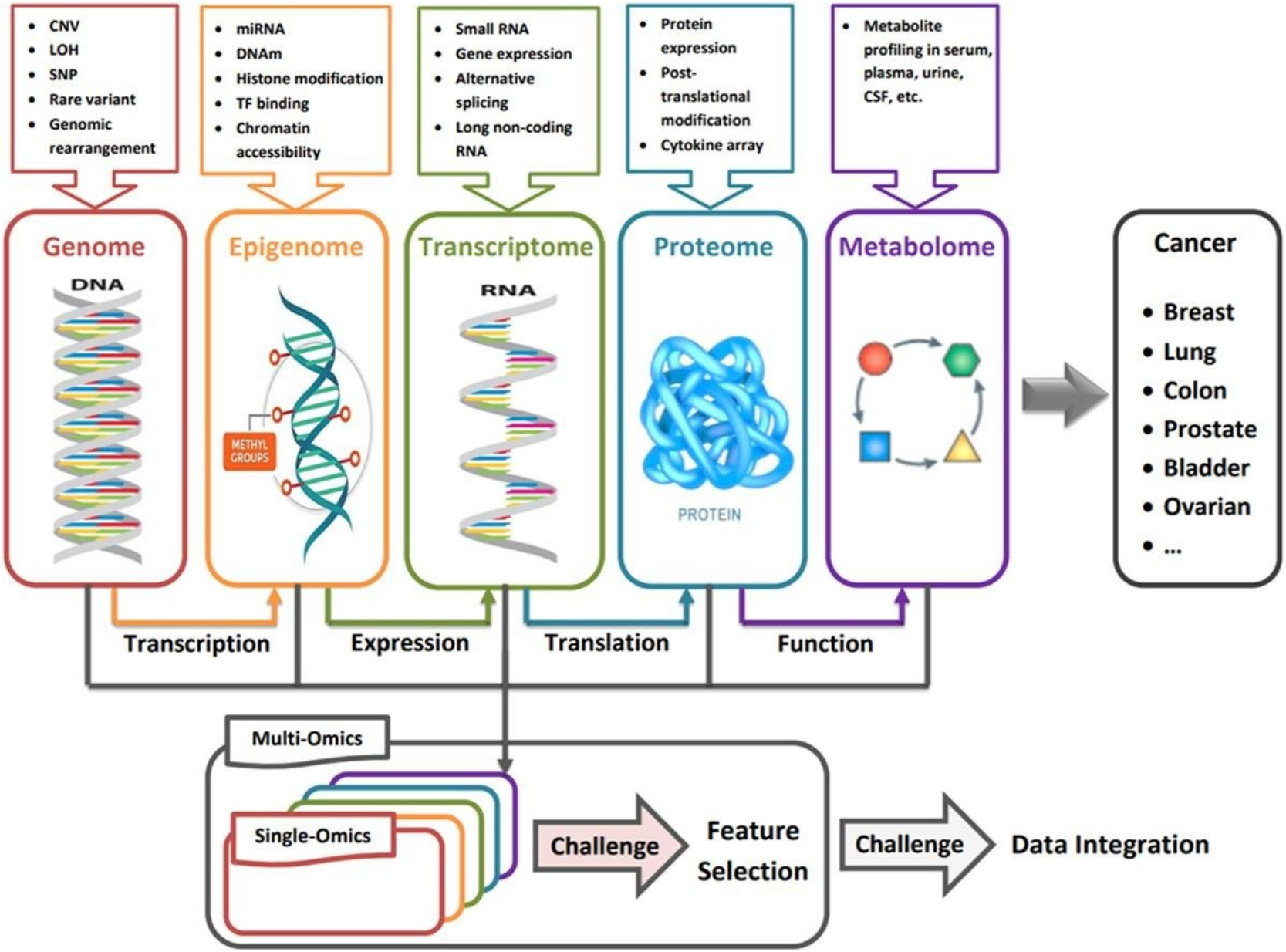
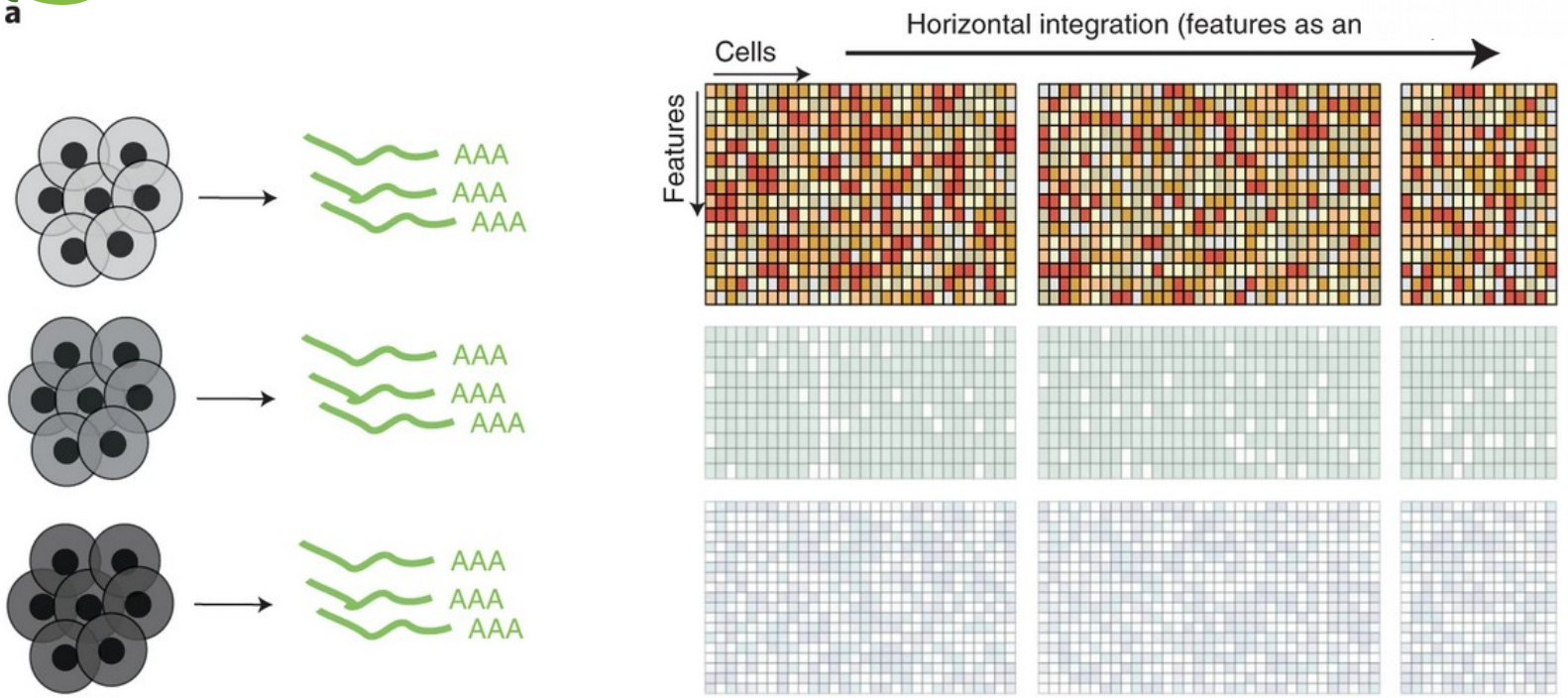
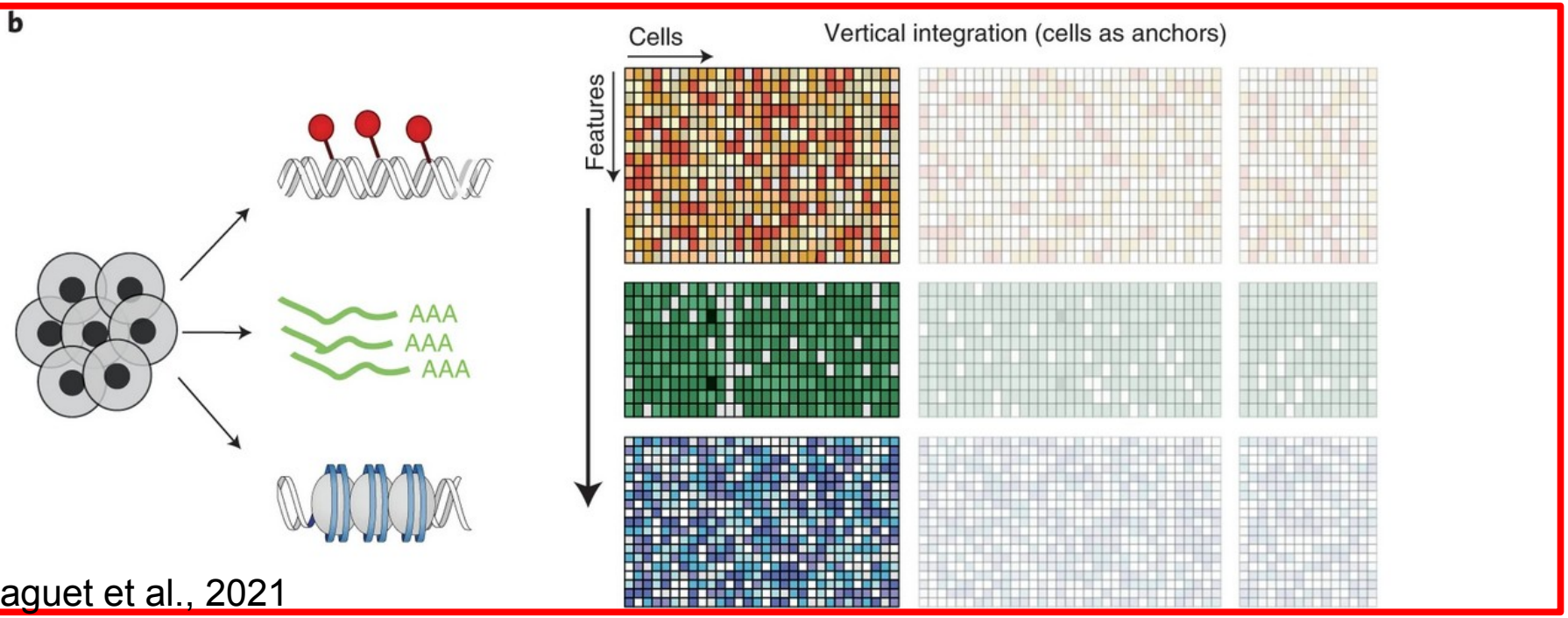


Image adapted from Momeni et al., J Biomed Inform. 2020

Integration Across Features vs. Samples



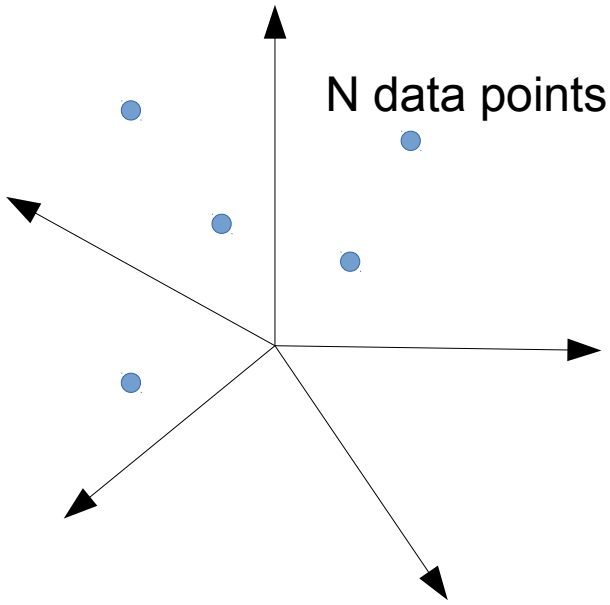


Argelaguet et al., 2021

Biological Data: High-Dimensional (P >> N)

Statistical observations:
e.g.samples, cells etc.

Features: genes, proteins,
microbes, metabolites etc.



N

P_1

0	3	1	0	2	3	8	1	1	3
1	1	0	0	7	1	2	2	3	3
1	2	2	0	0	6	7	1	2	2
1	2	3	10	0	4	6	1	0	5
3	2	2	1	4	3	2	1	6	0
7	4	4	5	3	9	6	1	6	1
7	1	1	5	2	8	9	1	3	6
5	0	1	6	2	0	0	0	1	5
1	6	3	3	4	6	2	0	1	1
1	2	2	4	1	1	3	0	8	2

OMIC1

$P_1 + P_2 + P_3 >>> N$

N

P_2

0	3	1	0	2	3	8	1	1	3
1	1	0	0	7	1	2	2	3	3
1	2	2	0	0	6	7	1	2	2
1	2	3	10	0	4	6	1	0	5
3	2	2	1	4	3	2	1	6	0
7	4	4	5	3	9	6	1	6	1
7	1	1	5	2	8	9	1	3	6
5	0	1	6	2	0	0	0	1	5
1	6	3	3	4	6	2	0	1	1
1	2	2	4	1	1	3	0	8	2

OMIC2

N

P_3

0	3	1	0	2	3	8	1	1	3
1	1	0	0	7	1	2	2	3	3
1	2	2	0	0	6	7	1	2	2
1	2	3	10	0	4	6	1	0	5
3	2	2	1	4	3	2	1	6	0
7	4	4	5	3	9	6	1	6	1
7	1	1	5	2	8	9	1	3	6
5	0	1	6	2	0	0	0	1	5
1	6	3	3	4	6	2	0	1	1
1	2	2	4	1	1	3	0	8	2

OMIC3

The Curse of Dimensionality complicates OMICs Integration

The Curse of Dimensionality

P is the number of features (genes, proteins, genetic variants etc.)
N is the number of observations (samples, cells, nucleotides etc.)

Biology / Biomedicine

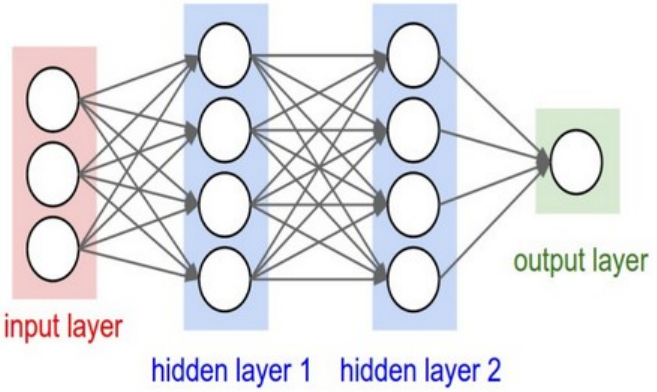
Bayesianism

P >> N

Frequentism

P ~ N

Deep Learning



P << N

Amount of Data

$$Y = \alpha + \beta X$$

$$\beta = (X^T X)^{-1} X^T Y$$

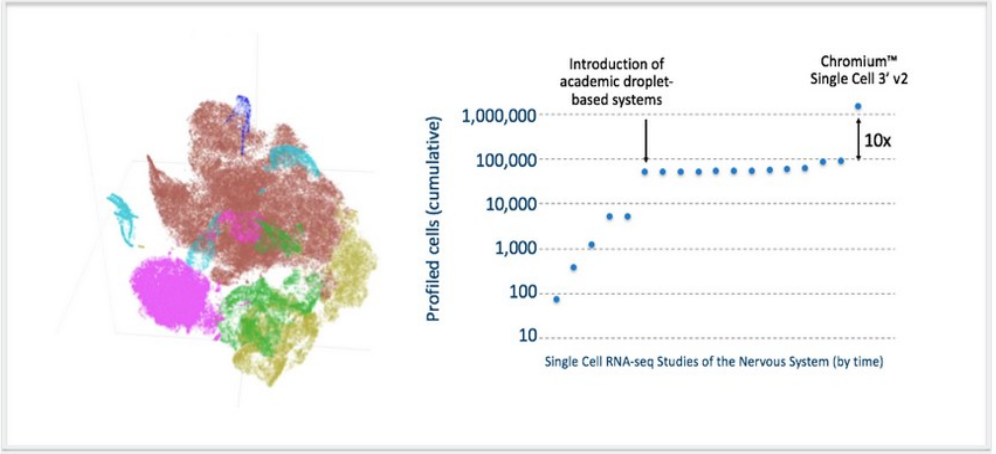
$$(X^T X)^{-1} \sim \frac{1}{\det(X^T X)} \cdots \to \infty, \quad n << p$$

Big Data in Single Cell

CAREERS BLOG 10X UNIVERSITY

10X GENOMICS SOLUTIONS & PRODUCTS RESEARCH & APPLICATIONS EDUCATION & RESOURC

« Back to Blog ‹ Newer Article Older Article ›



Our 1.3 million single cell dataset is ready to download

0 KUDOS

POSTED BY: grace-10x, on Feb 21, 2017 at 2:28 PM

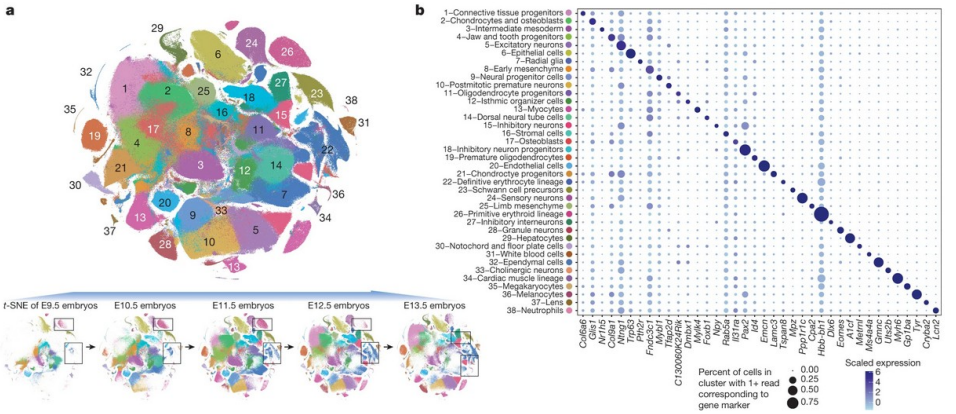
At ASHG last year, we announced our 1.3 Million Brain Cell Dataset, which is, to date, the largest dataset published in the single cell RNA-sequencing (scRNA-seq) field. Using the Chromium™ Single Cell 3' Solution (v2 Chemistry), we were able to sequence and profile 1,308,421 individual cells from embryonic mice brains. Read more in our application note Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium™ Single Cell 3' Solution.

Watch out Underfitting!
Paradise for Deep Learning!

MENU nature

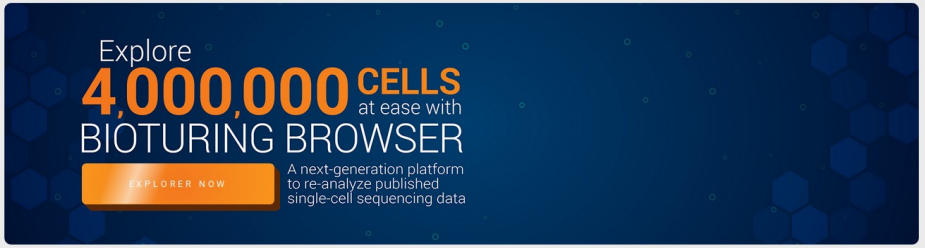
Fig. 2: Identifying the major cell types of mouse organogenesis.

From: The single-cell transcriptional landscape of mammalian organogenesis



a, t-SNE visualization of 2,026,641 mouse embryo cells (after removing a putative doublet cluster), coloured by cluster identity (ID) from Louvain clustering (in **b**), and annotated on the basis of marker genes. The same t-SNE is plotted below, showing only cells from each stage (cell numbers from left to right: n = 151,000 for E9.5; 370,279 for E10.5; 602,784 for E11.5; 468,088 for E12.5; 434,490 for E13.5). Primitive erythroid (transient) and definitive erythroid (expanding) clusters are boxed. **b**, Dot plot showing expression of one selected marker gene per cell type. The size of the dot encodes the percentage of cells within a cell type in

BioTuring Solutions Resources



Single Cell Analysis

5,500,000 cells will be indexed into BioTuring Single-cell Data Repository this September

by biomembers · August 30, 2019

Human Cell Atlas, single-cell data

We are glad to announce that we will upsize the current single-cell database in BioTuring Single-cell Browser to 5,500,000 cells this September. With this release, we will double the current number of publications indexed in BioTuring Single-cell Browser, and cross the number of cells hosted on available public single-cell data repositories like Human Cell Atlas (HCA) and Broad Institute's Single-cell Portal.

Search

RECENT POSTS

A new tool to interactively visualize single-cell objects (Seurat, Scanpy, SingleCellExperiments, ...)
September 26, 2019

5,500,000 cells will be indexed into BioTuring Single-cell Data Repository this September
August 30, 2019

How to define and evaluate OMICs Integration?

What I Mean by OMICs Integration

mix Omics

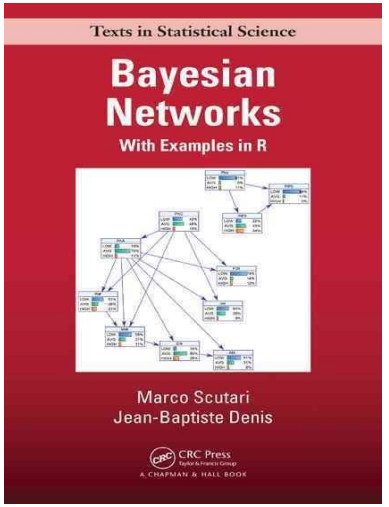
Exploration and Integration of Omics datasets

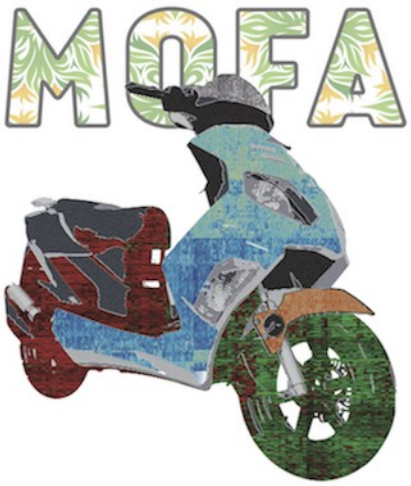
OnPLS

JIVE

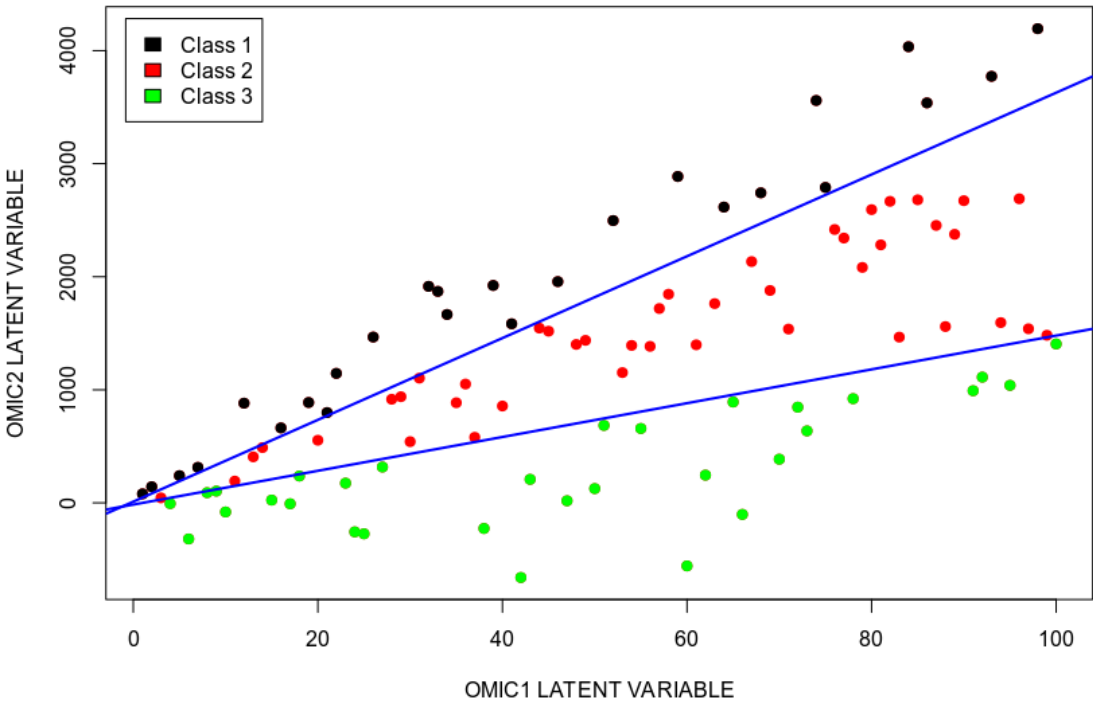
DISCO

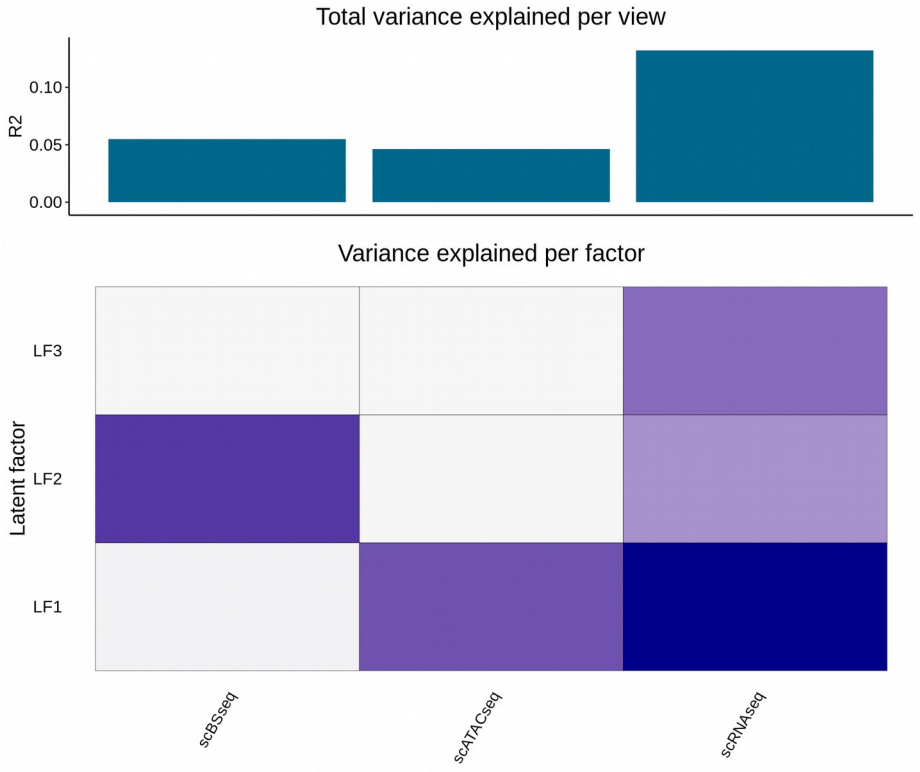
Clustering of Clusters





Idea Behind OMICs Integration: See Patterns Hidden in Individual OMICS





How I Evaluate OMICs Integration, Data Science: Boost in Prediction

TEXT (78%)

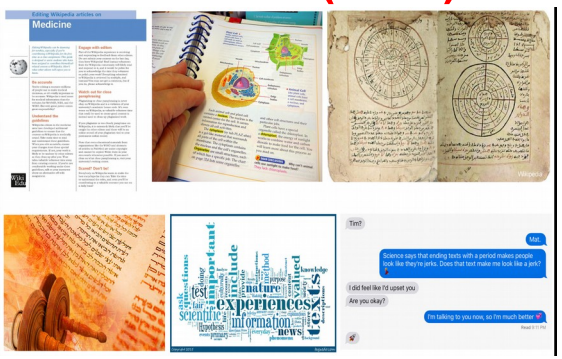
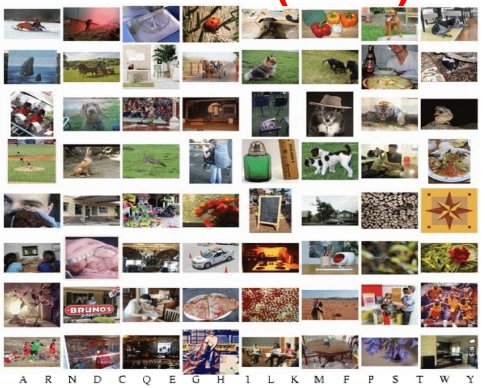
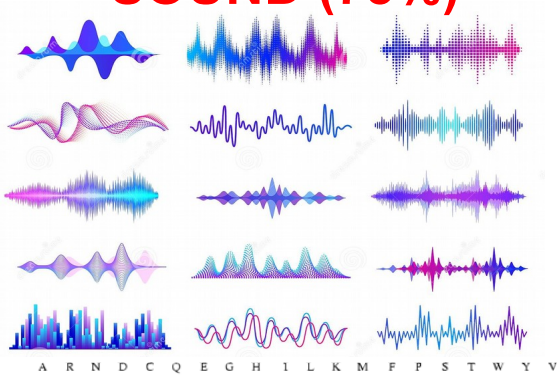
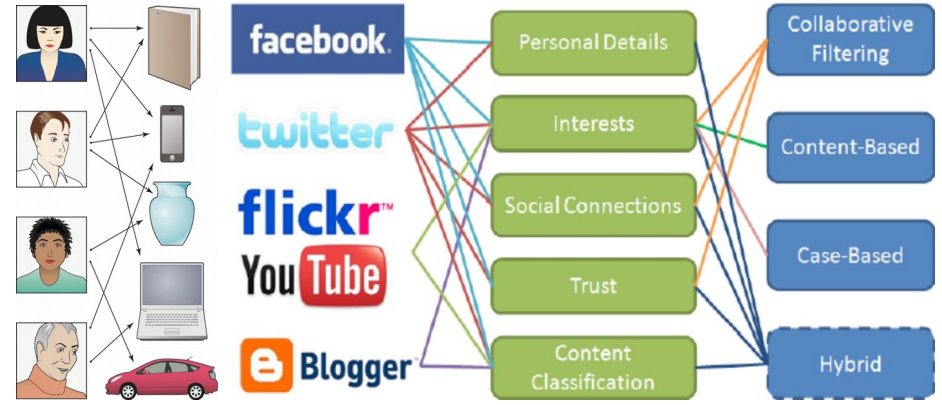


IMAGE (83%)



SOUND (75%)



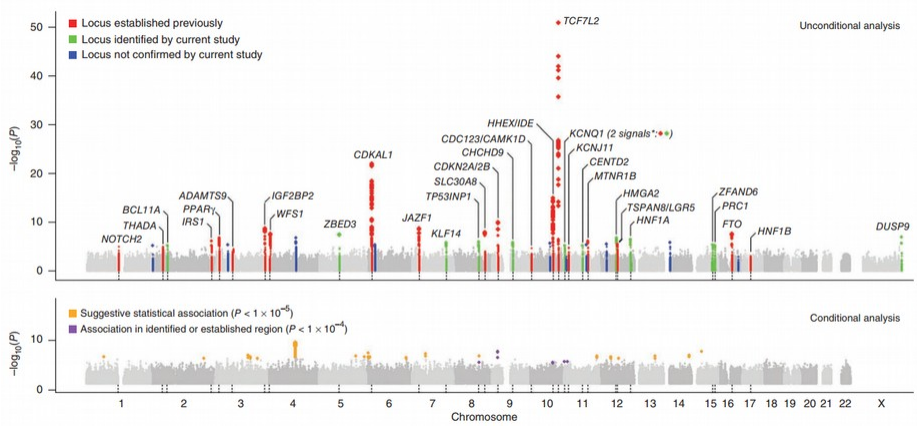


Predict Facebook user interests

Data Integration Accuracy: 96%

Prediction is an Ultimate Criterion of Successful OMICS Integration

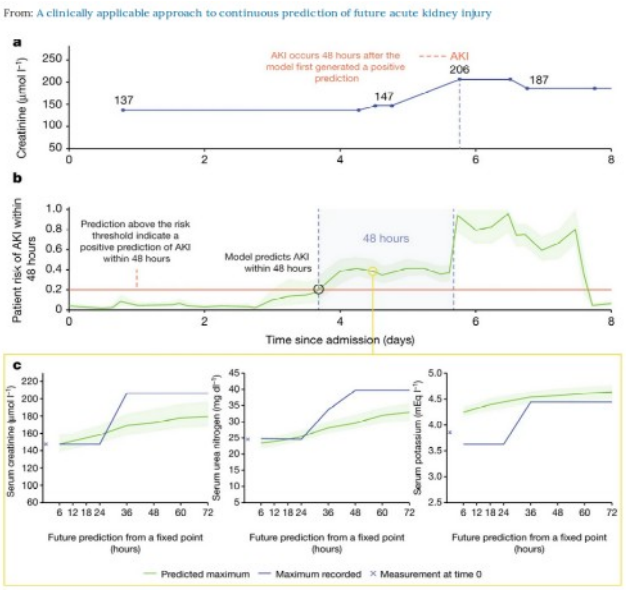
Statistics searches for candidates



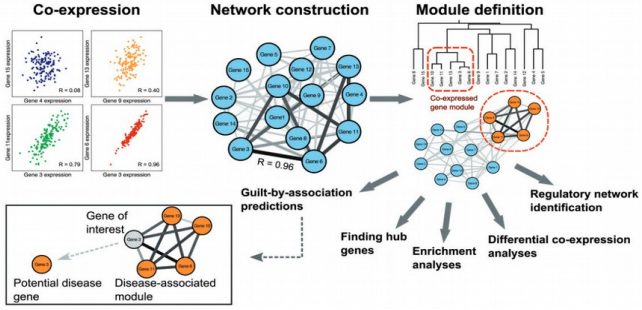
Consequence

B. Maher, Nature 456, 18-21 (2008)

Machine Learning optimizes prediction

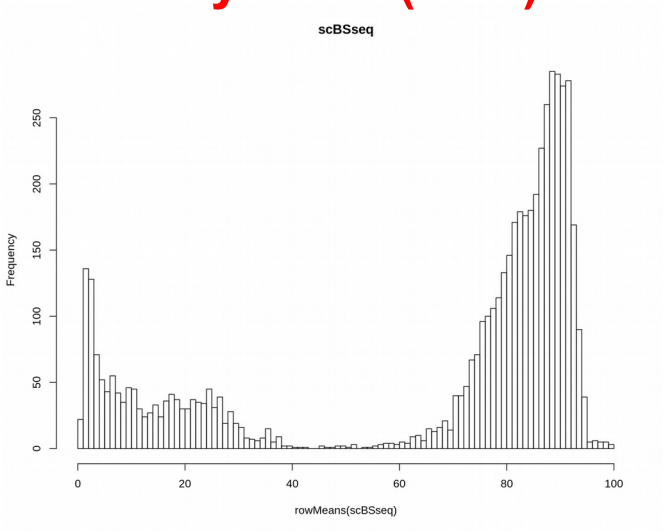


Consequence

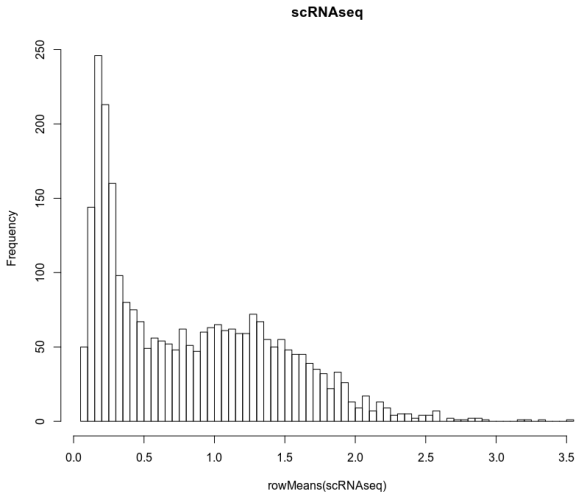


How I Evaluate OMICs Integration, Life Sciences: Boost in Prediction

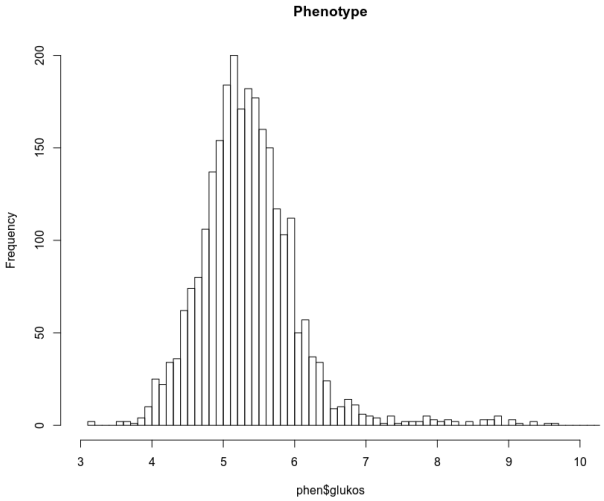
Methylation (78%)



Gene Expression (83%)



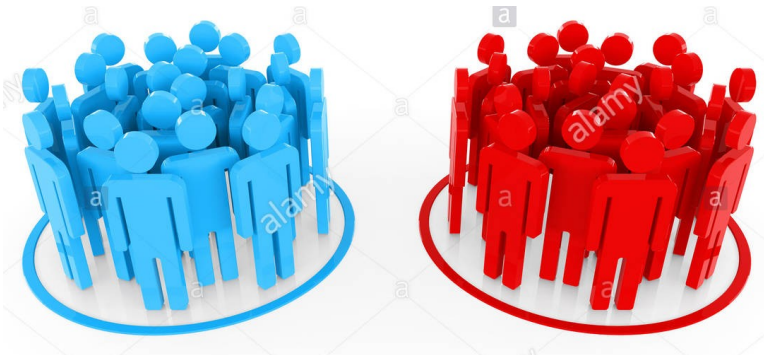
Phenotype (75%)



1) Convert to common space:
Neural Networks, SNF, UMAP

2) Explicitly model distributions:
MOFA, Bayesian Networks

3) Extract common variation:
PLS, CCA, Factor Analysis



HEALTHY **SICK**

Data Integration Accuracy: 96%

What Method of Integration to Select?

	Linear	Non-Linear
Supervised	PLS / OPLS / mixOmics, LASSO / Ridge / Elastic Net	Neural Networks, Random Forest, Bayesian Networks
Unsupervised	Factor Analysis / MOFA	Autoencoder, SNF, UMAP, Clustering of Clusters

For Example:

1) With ~100 samples it is a good idea to do **linear** OMICs integration
2) T2D is a phenotype of interest, therefore **supervised** integration

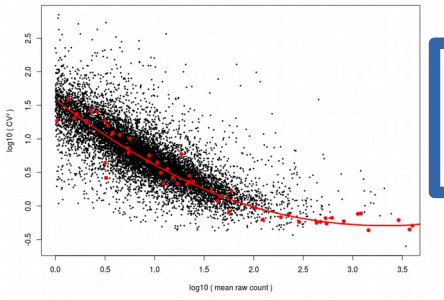
Workflow of OMICs Integration

Data Set (4 OMICs)

Train Set (n = 80)

Test Set (n = 20)

unsupervised



Feature Pre-Selection

supervised

Evaluation

OMICs Integration

Trained Model

Protocol of OMICs Integration

1) Check that there is a relation between the OMICs (MOFA)

2) Choose integrative model based on amount of data and goal (linear, supervised)

3) Do feature pre-selection (supervised or unsupervised) on train data set

4) Integrate the OMICs using your favorite model chosen in 2) on train data set

5) Compare prediction of integrative model with predictions from individual OMICs

National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

LUNDS UNIVERSITET