ICGEB TRIESTE
ITALY

ICGEB NEW DELHI
INDIA

ICGEB CAPE TOWN
SOUTH AFRICA

ICGEB International Centre for Genetic Engineering and Biotechnology | Science for Development

An Intergovernmental Organisation for research, training and technology transfer in Life Sciences to promote sustainable global developments

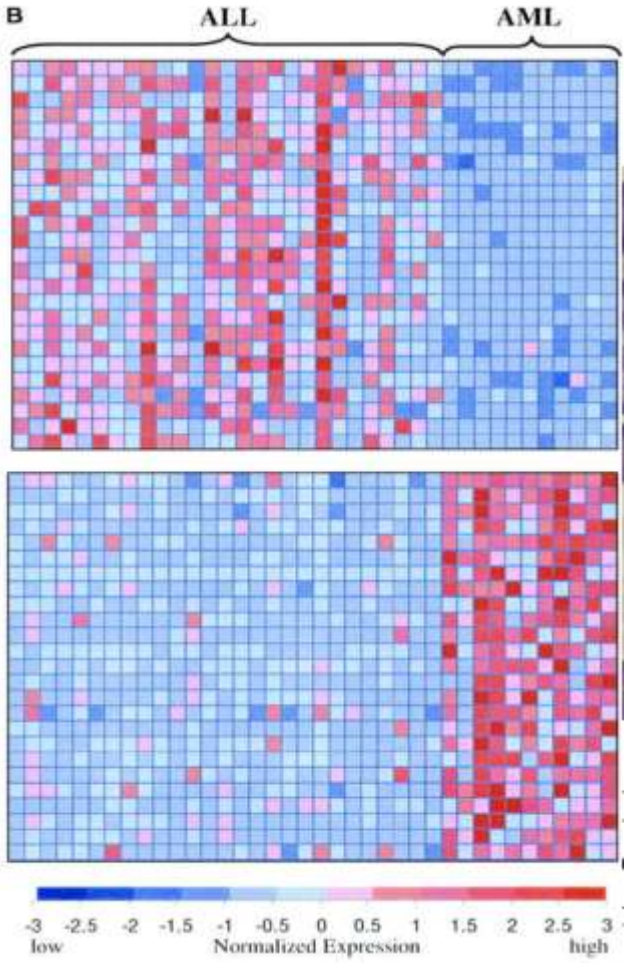
Artificial Intelligence for Single-Cell Genomics Sequencing Data Analysis

Stefano Cacciatore

13 November 2025, New Delhi

www.icgeb.org

Beginning of AI revolution



Science

Current Issue First release papers Archive About Submit manuscript

REPORTS

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. GOLUB, D. K. SLONIM, P. TAMAYO, C. HUARD, M. GAASENBEEK, J. P. MESIROV, H. COLLER, M. L. LOH, J. R. DOWNING, [...] E. S. LANDER +3 authors Authors Info & Affiliations

SCIENCE · 15 Oct 1999 · Vol 286, Issue 5439 · pp. 531-537 · DOI: 10.1126/science.286.5439.531

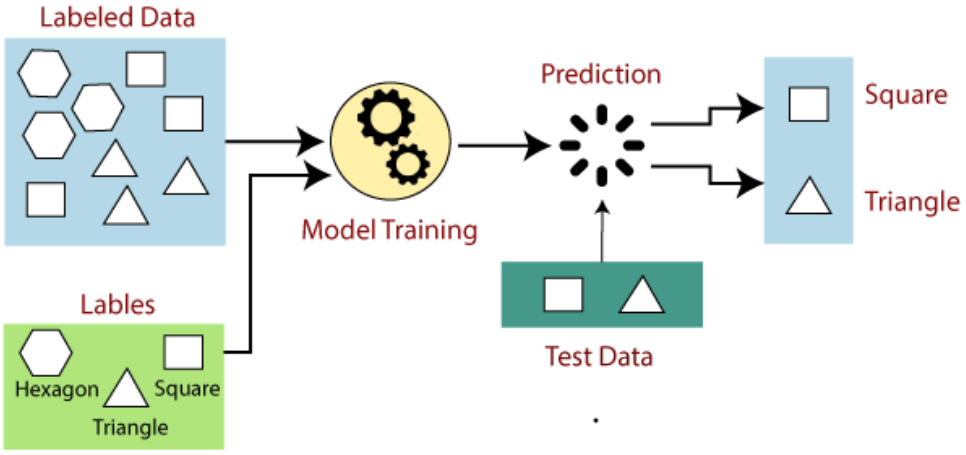
A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes.

Machine Learning

Supervised learning is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time.

Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention (hence, they are "unsupervised").

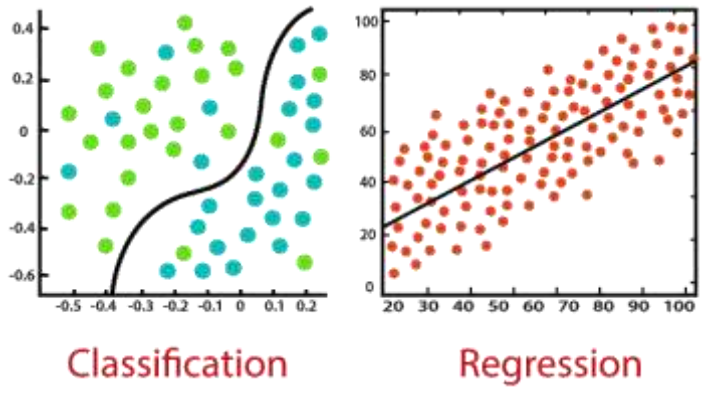
Supervised Learning



Classification and Regression

Regression and Classification algorithms are Supervised Learning algorithms. Both the algorithms are used for prediction in Machine learning and work with the labeled datasets. But the difference between both is how they are used for different machine learning problems.

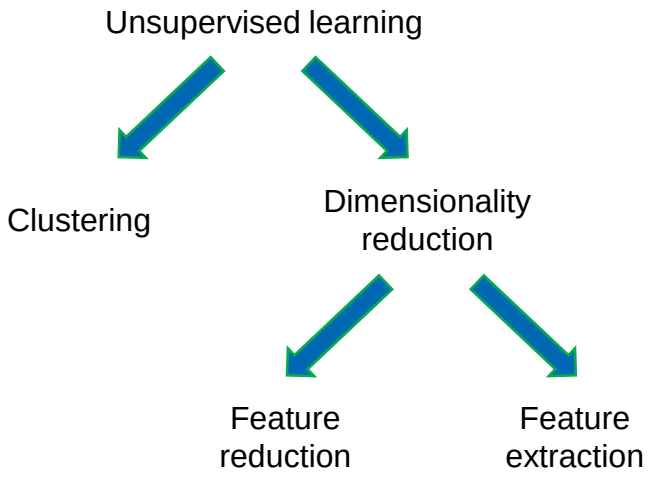
The main difference between Regression and Classification algorithms that Regression algorithms are used to predict the continuous values such as age and BMI. and Classification algorithms are used to predict/Classify the discrete values such as disease severity.



Supervised learning is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time.

Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention (hence, they are "unsupervised").

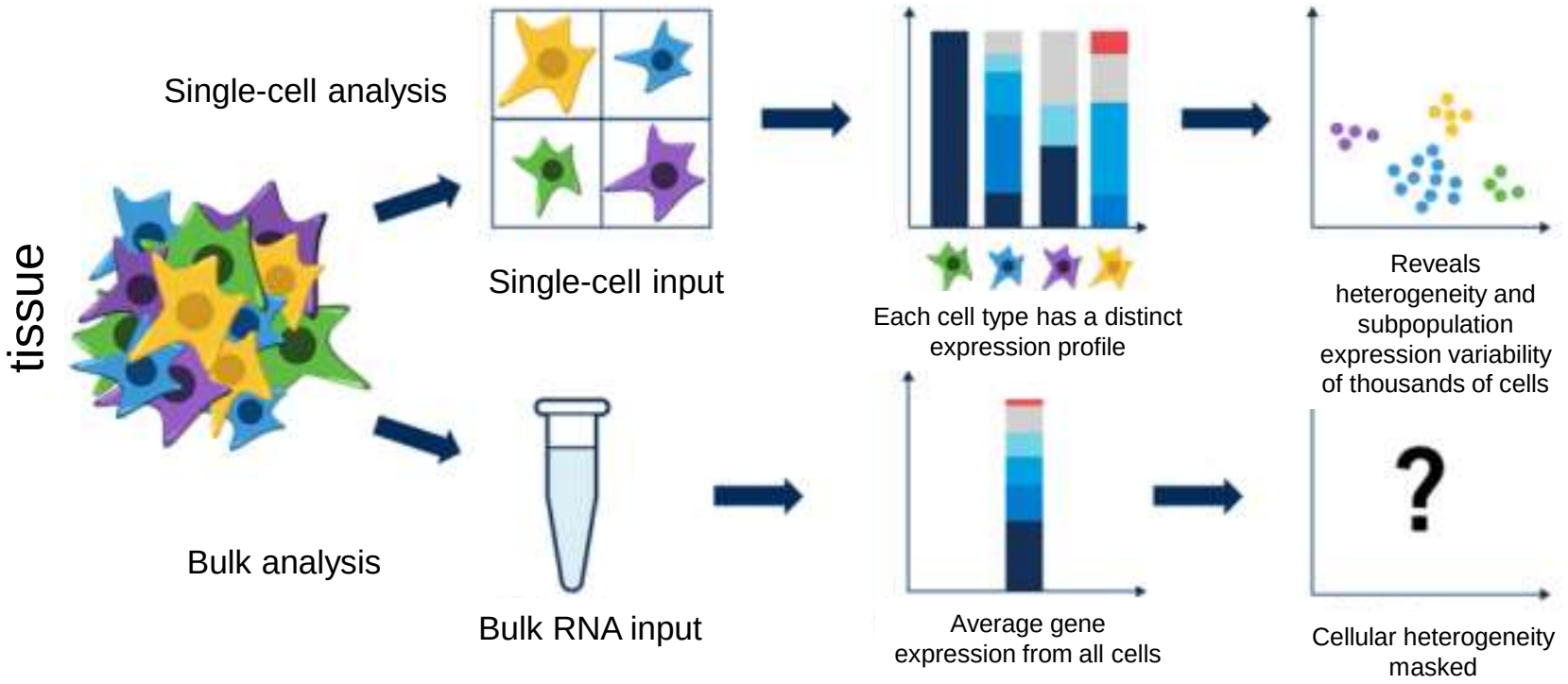
Unsupervised Learning



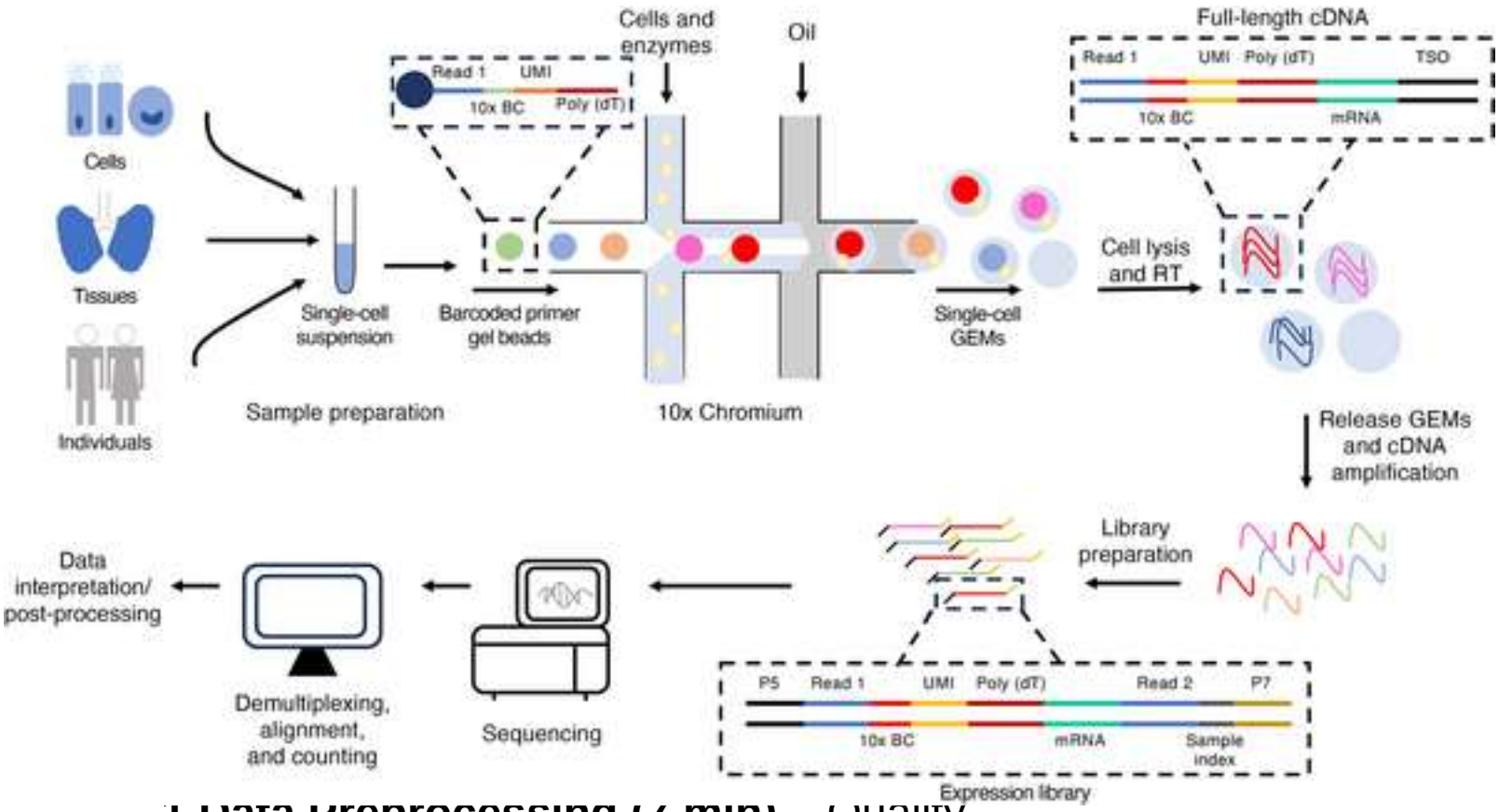
Clustering is a data mining technique for grouping unlabeled data based on their similarities or differences. For example, K-means clustering algorithms assign similar data points into groups, where the K value represents the size of the grouping and granularity. This technique is helpful for market segmentation, image compression, etc.

Dimensionality reduction is a learning technique used when the number of features (or dimensions) in a given dataset is too high. It reduces the number of data inputs to a manageable size while also preserving the data integrity. Often, this technique is used in the preprocessing data stage, such as when autoencoders remove noise from visual data to improve picture quality.

Single-cell transcriptomics



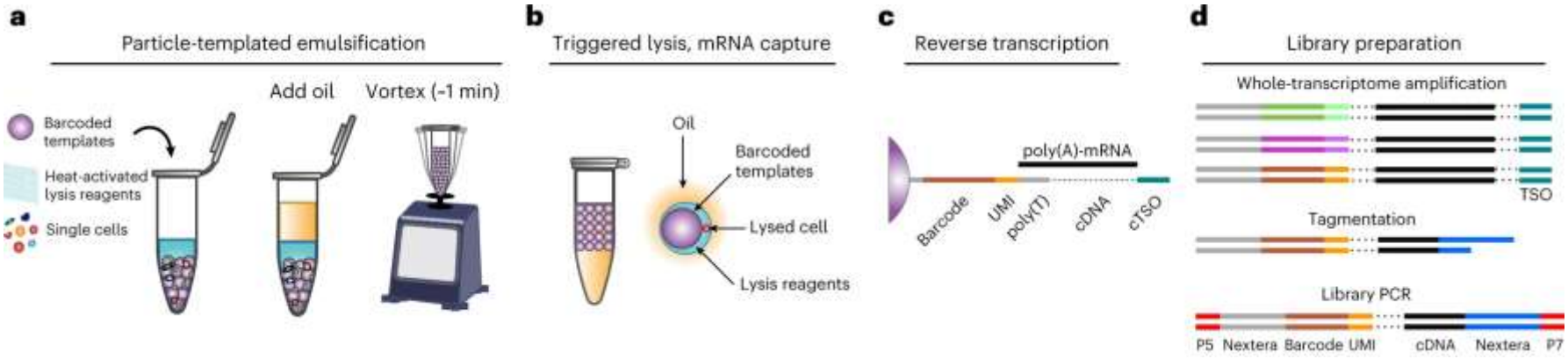
Single-cell transcriptomics workflow: Chromium technologies



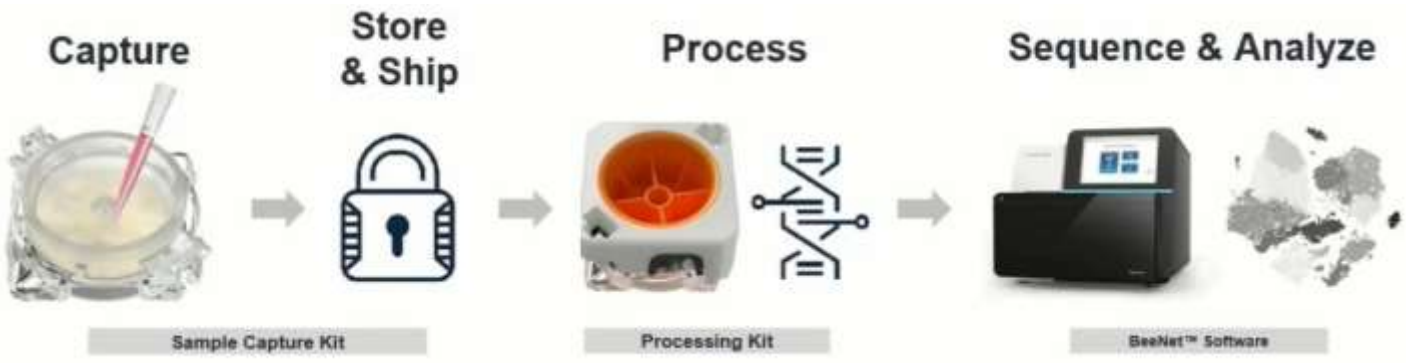
1.Data Preprocessing (7 min) – Quality control, normalization, feature selection, and batch correction

Single-cell transcriptomics workflow: instrument-free

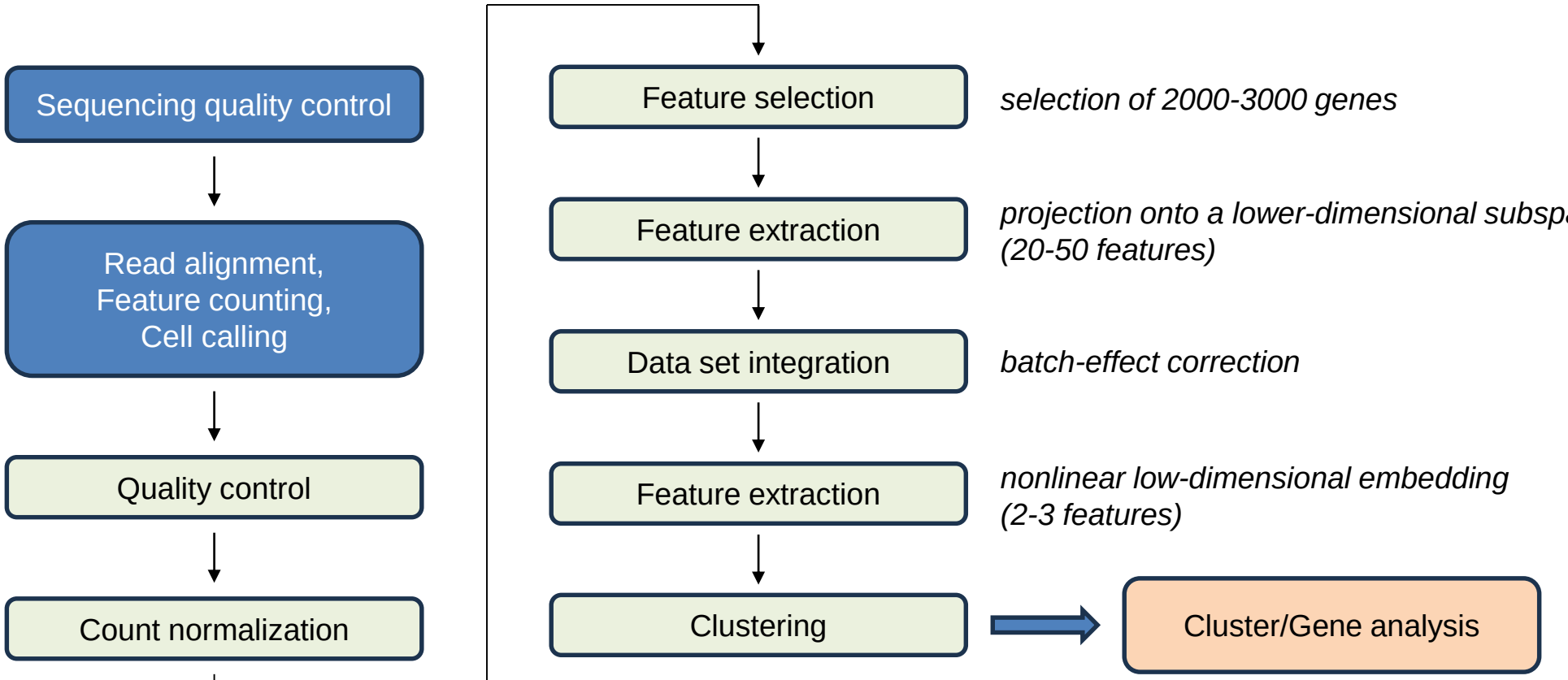
Fluent BioScience



Honeycomb



Single-cell: standard pre-processing workflow



Quality control: Not every droplet is useful

A single happy cell in a droplet is ideal

- Complex transcriptome
- Average number of genes detected

Empty droplet: No cell in a droplet

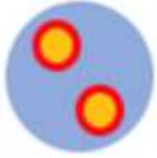
- No genes detected

Droplet with ambient RNA

- Low complex transcriptome
- Genes detected much lower than average genes per cell

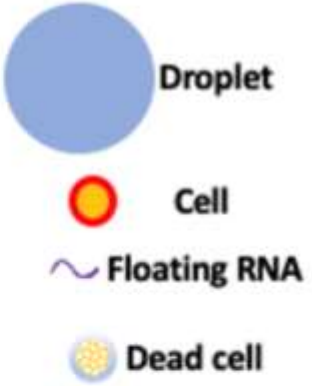
Droplet with dead cell

- Enriched for mitochondrial genes



Droplet with multiple cell

- Very complex transcriptome
- Genes detected much higher than average genes per cell



Quality control: Not every droplet is useful

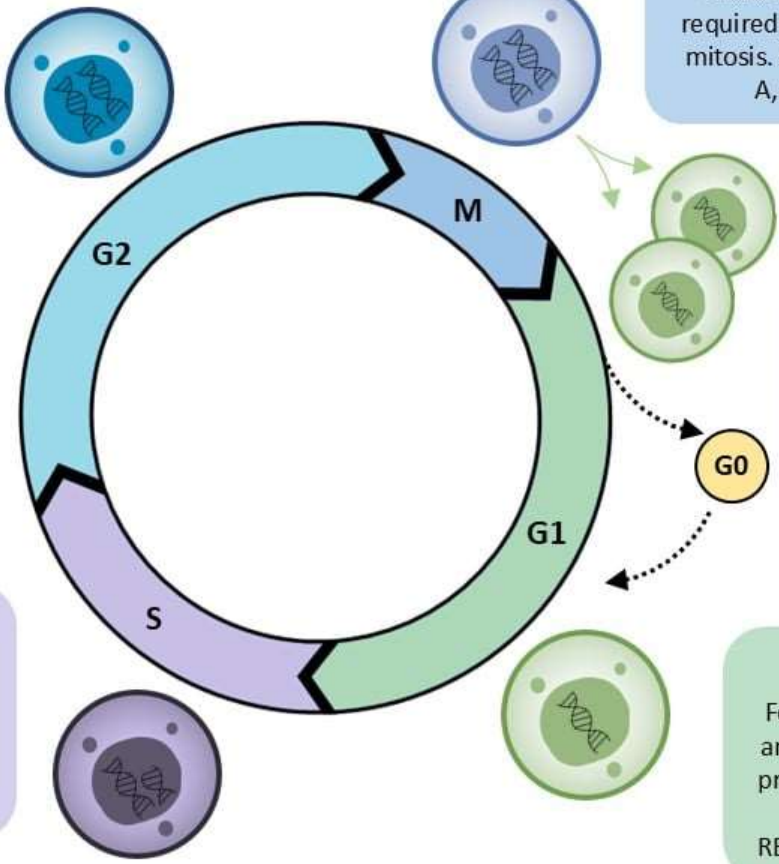
Aim of QC is …

- To remove undetected genes
- To remove empty droplets
- To remove droplets with dead cells
- To remove Doublet/multiplet
- Ultimately To filter the data to only include true cells that are of high quality

Above is achieved by applying hard cut-off or adaptive cut-off on …

- Number of genes detected per cell
- Percent of mitochondrial genes per cell
- Number of UMIs/transcripts detected per cell

Quality control: Not every droplet is useful



Gap 2 (G2)
After DNA synthesis, G2 serves as a preparatory phase before mitosis. Key proteins expressed in G2 include CDK1, cyclins A, B and F, MYT1, WEE1, PLK1, and FOXM1.

Mitosis (M)
There are several notable proteins that are required to initiate and complete the stages of mitosis. Among these are APC/C, CDK1, cyclin A, CDH1, AURKB, and Histone H3.

DNA content
The amount of cellular DNA, which is doubled in G2/M relative to G1/G0, can be used to distinguish cell cycle stages.

Gap 0 (G0)
From the G1 phase, quiescent, senescent, or differentiated cells can enter G0. Diminished levels of certain proteins such as Ki-67 and FOXM1 are hallmarks of this phase.

DNA Synthesis (S)
DNA replication proteins—such as CDC45, MCMs, and PCNA—are activated and expressed in the S phase. Key proteins needed to transition into this phase include SKP2, CDK2, CDC25A and Cyclin E.

Gap 1 (G1)
Following cell division, cells in G1 grow and prepare for DNA synthesis. Notable proteins expressed in G1 include APC/C, CDH1, Cyclin D/E, CDK2/4/6, RB1, RBL1/2, p15, p16, p18, p19, and CIP/KIP.

Normalization

Purpose of normalization: Adjust raw single-cell expression values to make cells comparable, accounting for differences in sequencing depth and capture efficiency.

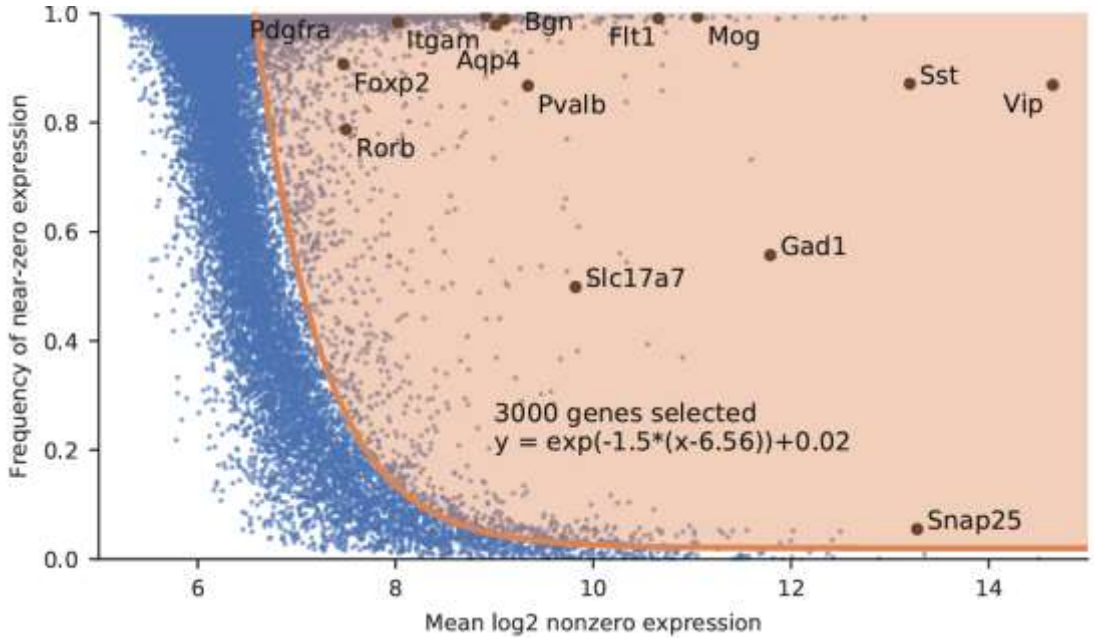
Common approach: Global-scaling methods divide each cell's counts by its total expression, multiply by a scale factor, and apply a log transformation to stabilize variance.

Alternative strategies: Model-based or variance-stabilizing methods (e.g., regularized negative binomial approaches) can better account for technical noise and varying RNA content across cells.

Variable selection

Goal: Identify genes that exhibit high variability across cell types

Focusing on highly variable genes reduces noise from ubiquitous or uninformative transcripts, improving downstream dimensionality reduction and clustering.

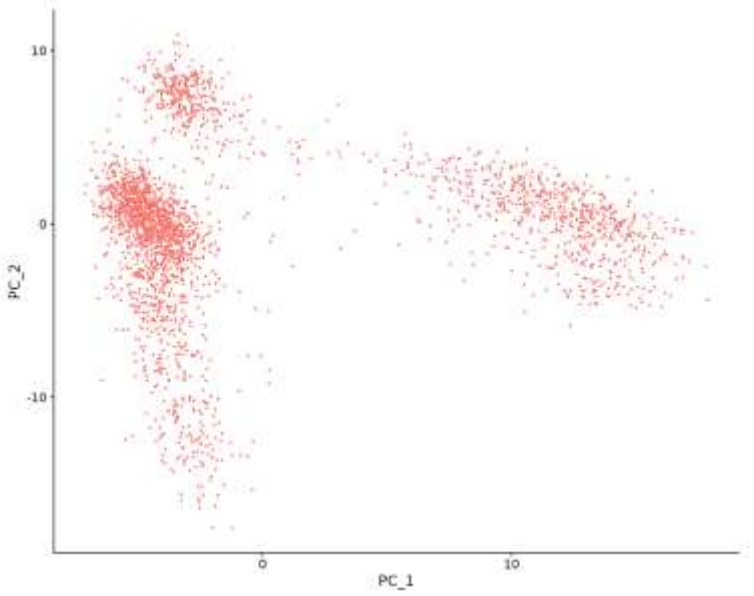


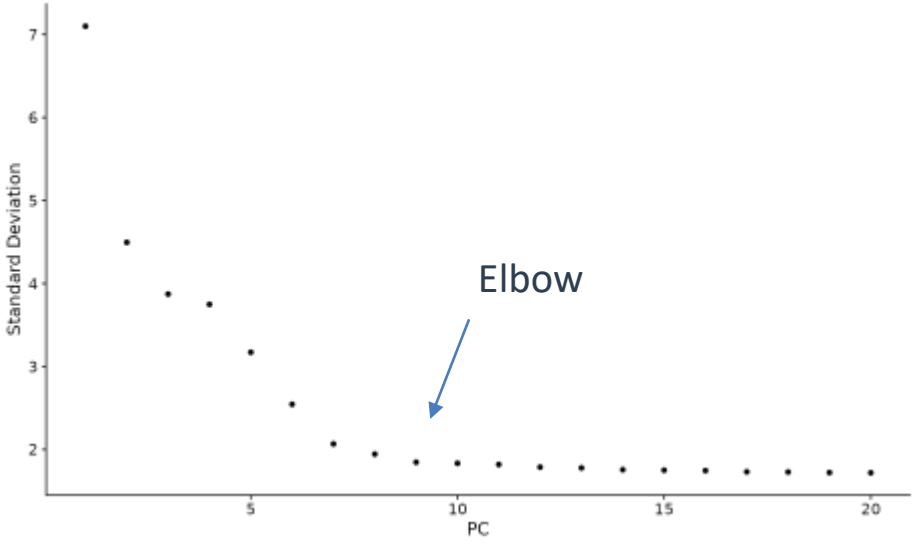
Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019 Nov 28;10(1):5416. doi: 10.1038/s41467-019-13056-x. PMID: 31780648; PMCID: PMC6882829.

Feature extraction: Principal Component Analysis

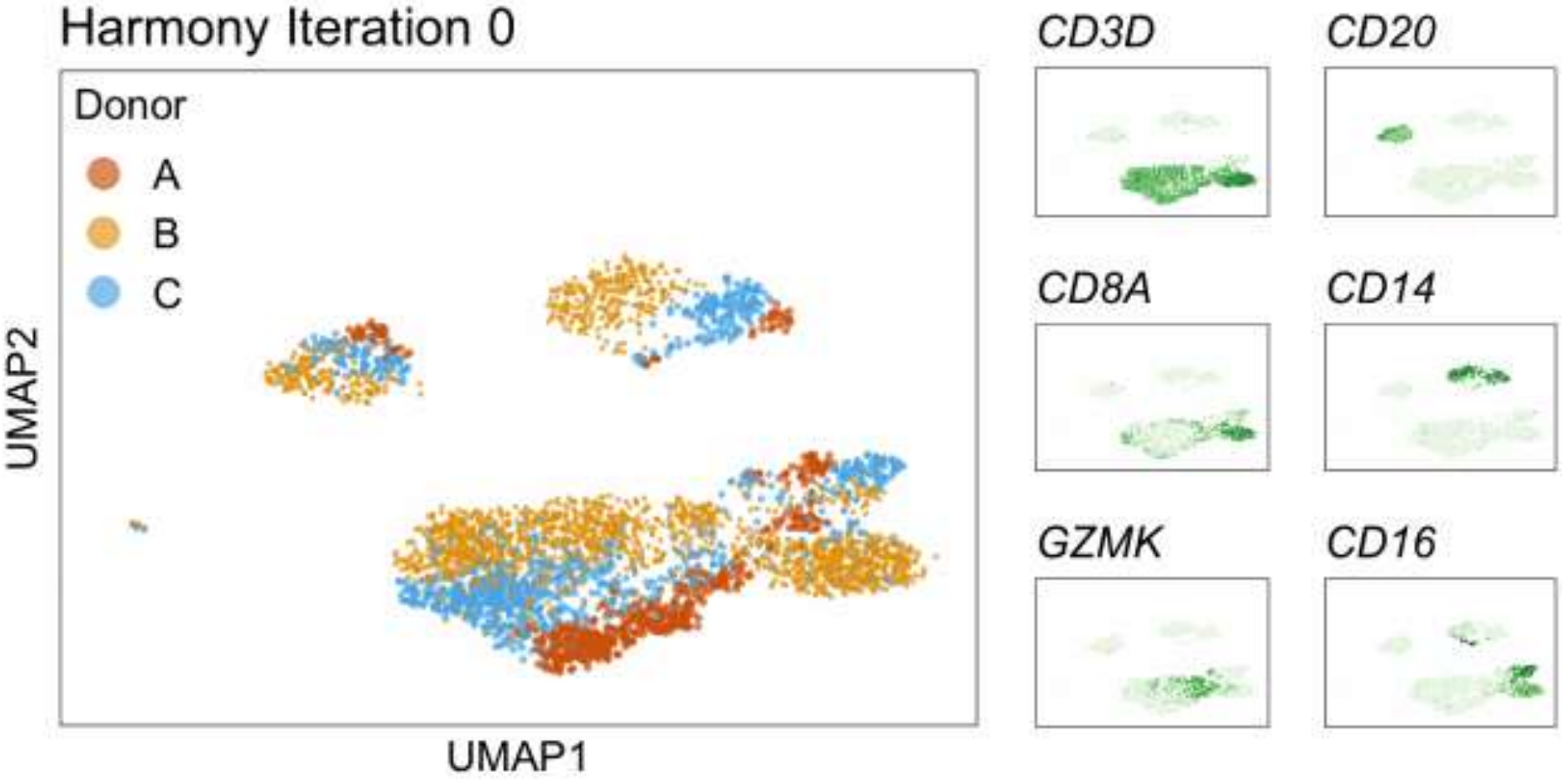
To overcome the extensive technical noise inherent in single-cell RNA-seq data, cells are analyzed based on their **principal component (PC) scores**, where each PC represents a **metafeature** summarizing correlated gene expression patterns.

The top principal components provide a **robust, low-dimensional representation** of the dataset. However, an important question remains: **how many components should be included 10, 20, or 100?**





Batch-effect correction



Feature extraction: nonlinear low-dimensional embedding

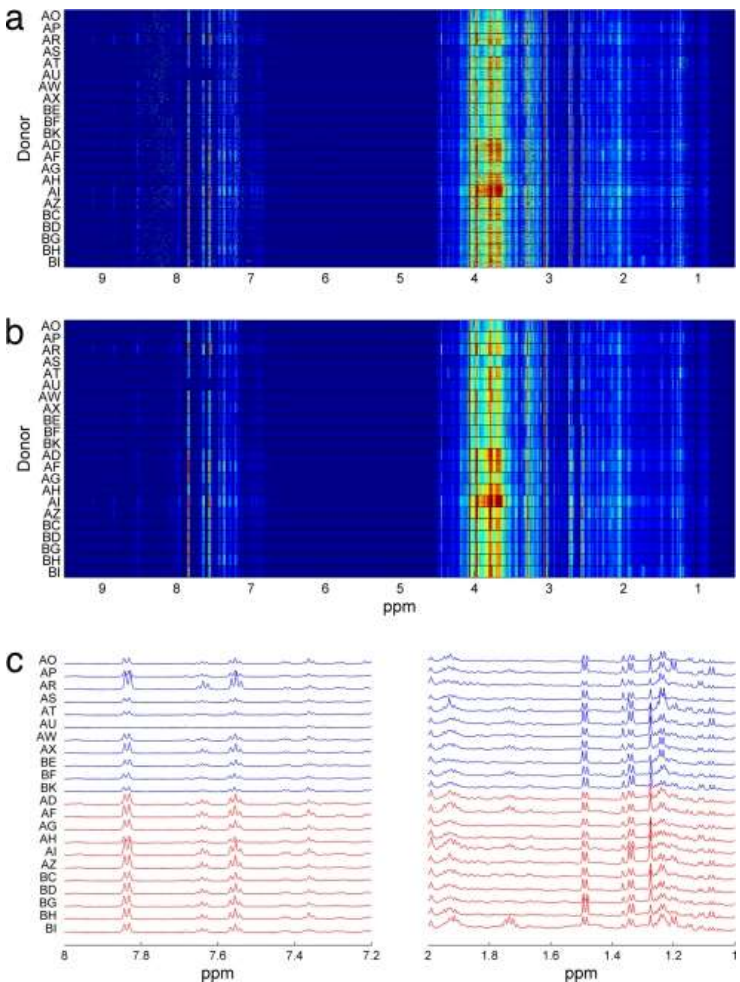
Purpose: Visualize high-dimensional single-cell data in 2D/3D while preserving local and global structure.

t-SNE: Focuses on maintaining local neighborhoods; good for revealing clusters but may distort global distances.

UMAP: Balances local and global relationships, faster and better preserves data topology and continuity.

Urine from Healthy Donors

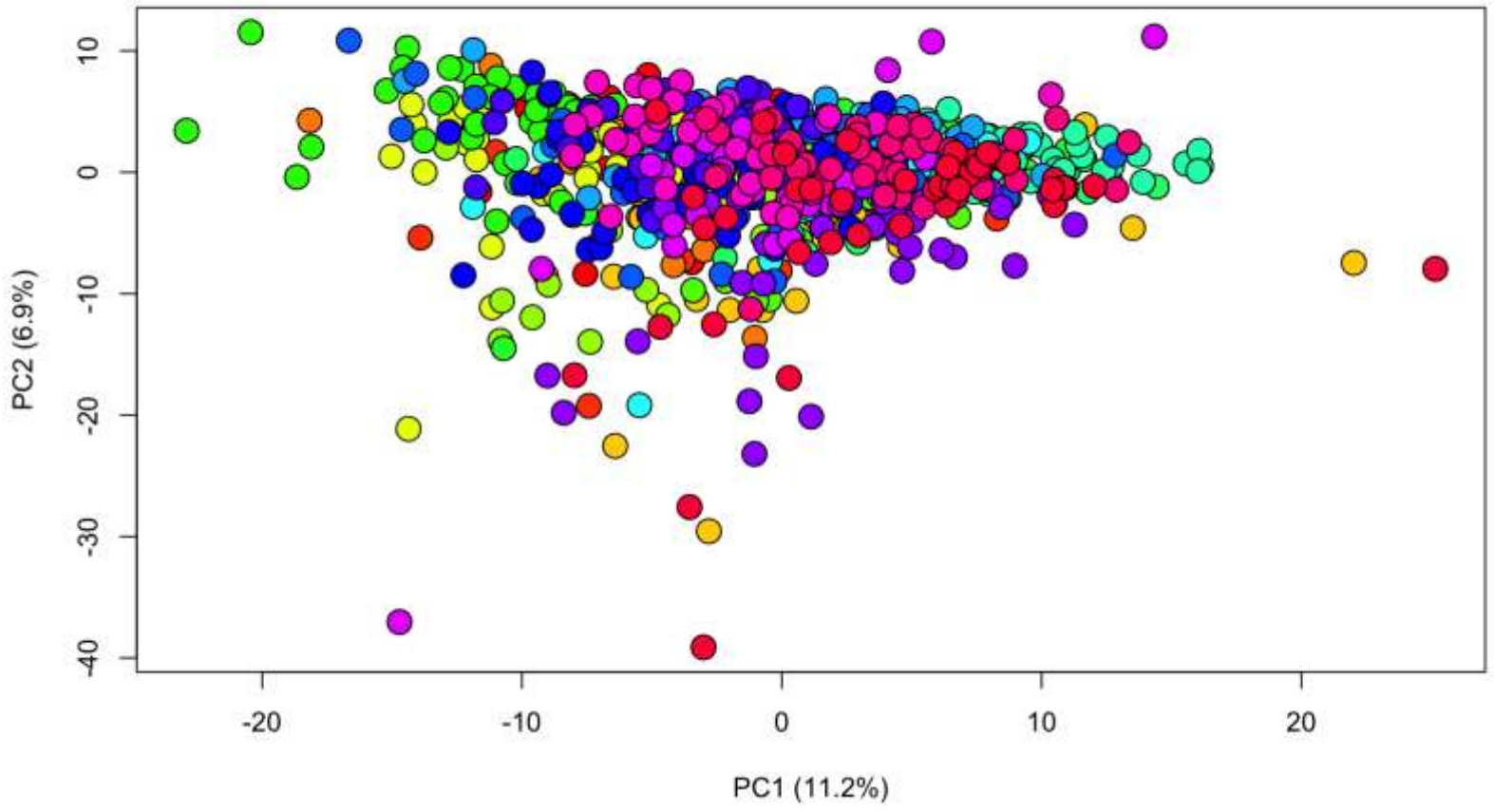
The data belong to a cohort of 22 healthy donors (11 male and 11 female) where each provided about 40 urine samples over the time course of approximately 2 months, for a total of 873 samples. Each sample was analysed by Nuclear Magnetic Resonance Spectroscopy. Each spectrum was divided in 450 spectral bins.



Proc Natl Acad Sci U S A. 2008 Feb 5;105(5):1420-4.
Evidence of different metabolic phenotypes in humans
Michael Assfalg, Ivano Bertini, Donato Colangiuli, Claudio Luchinat, Hartmut Schäfer, Birk Schütz, Manfred Spraul

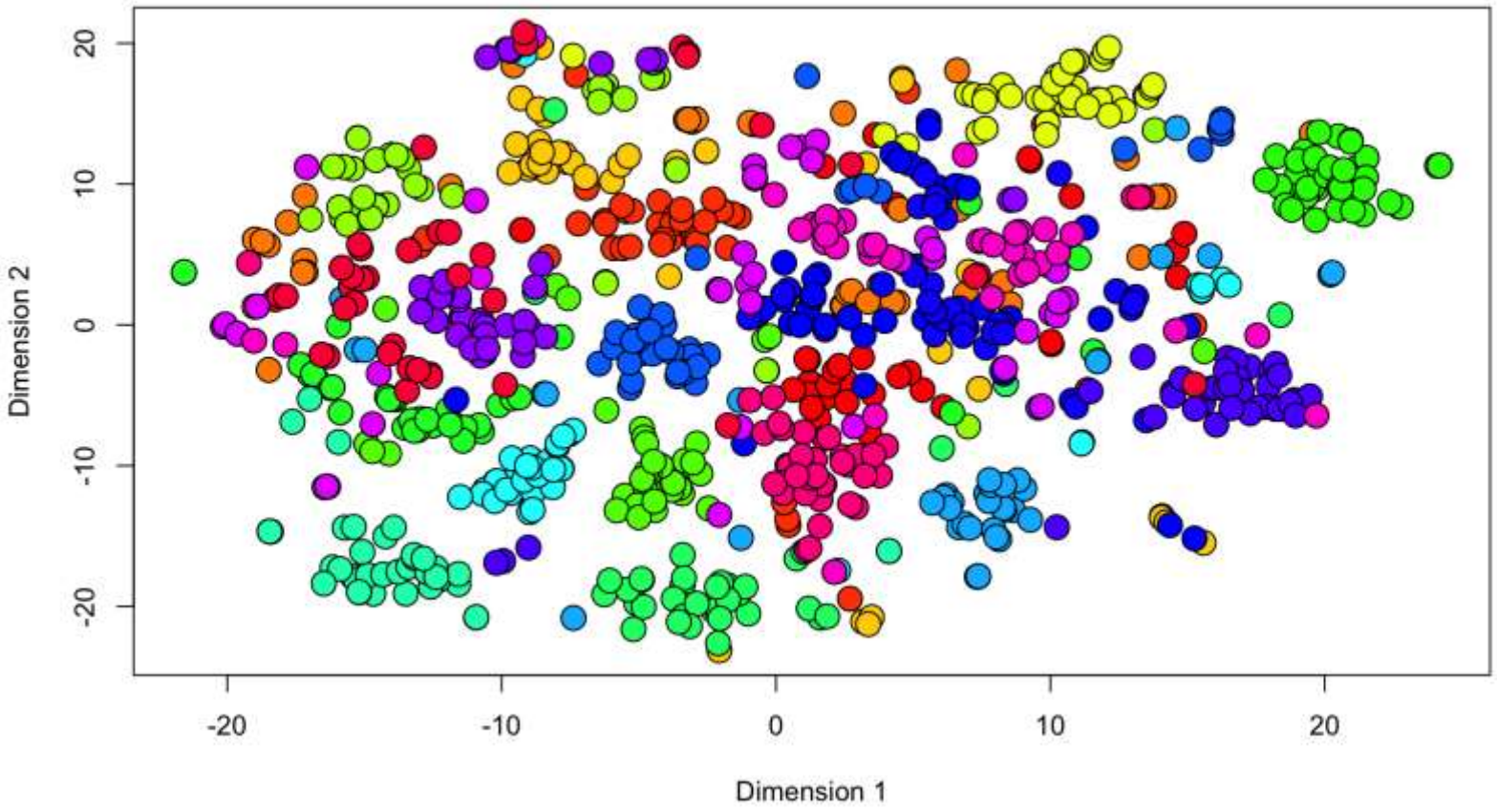
Urine from Healthy Donors

Principal Component Analysis



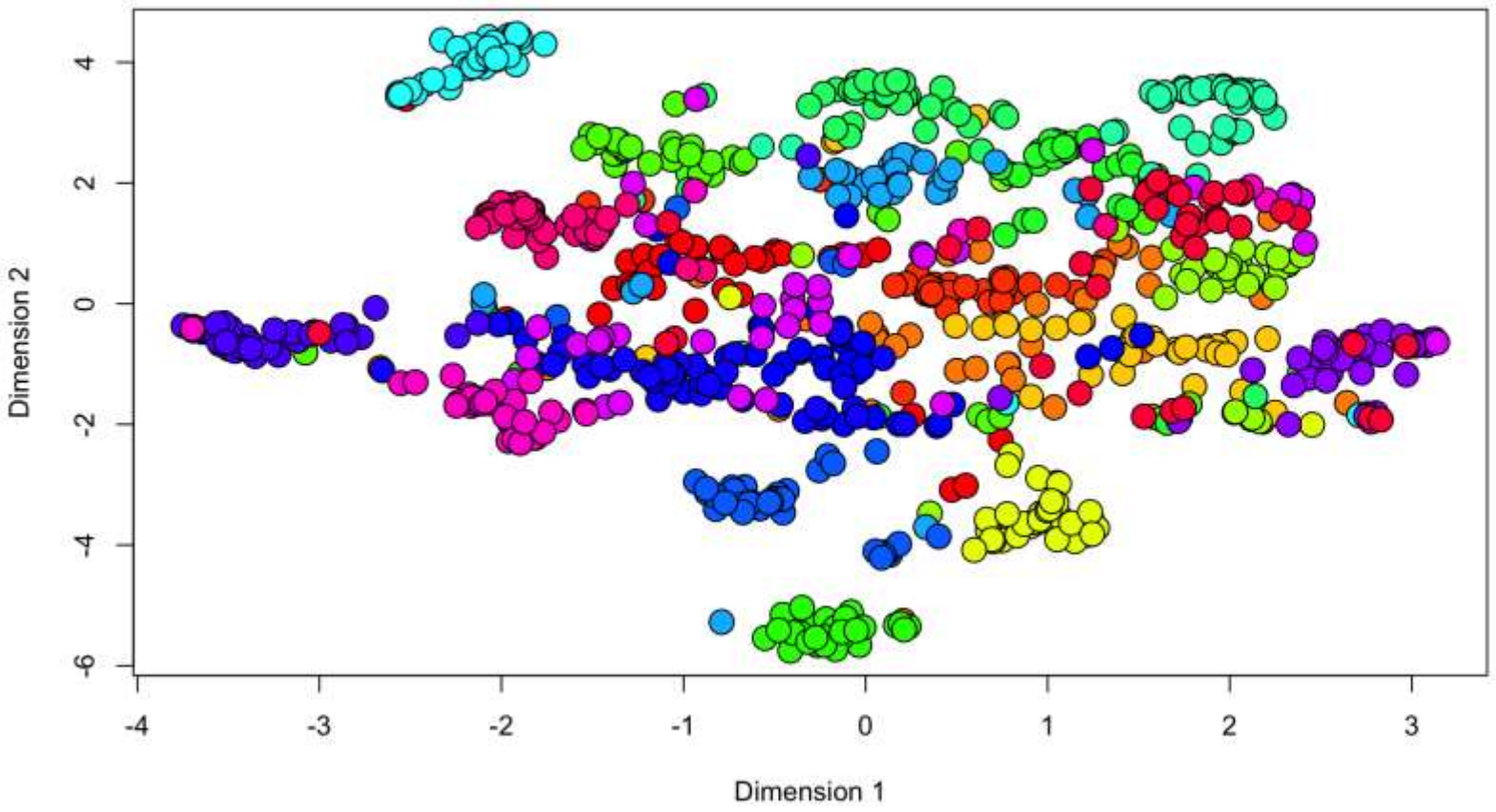
Urine from Healthy Donors

t-SNE



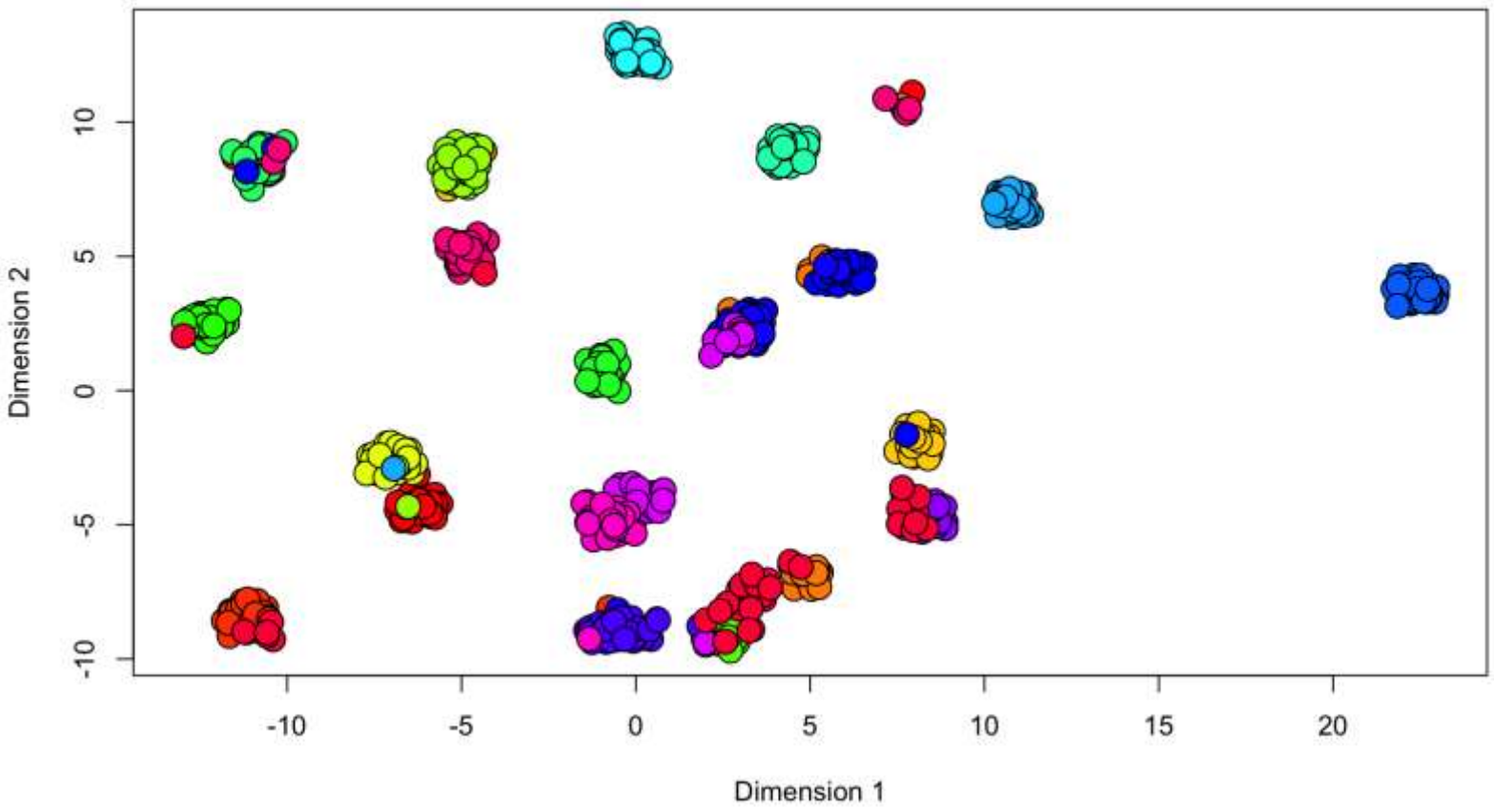
Urine from Healthy Donors

t-SNE



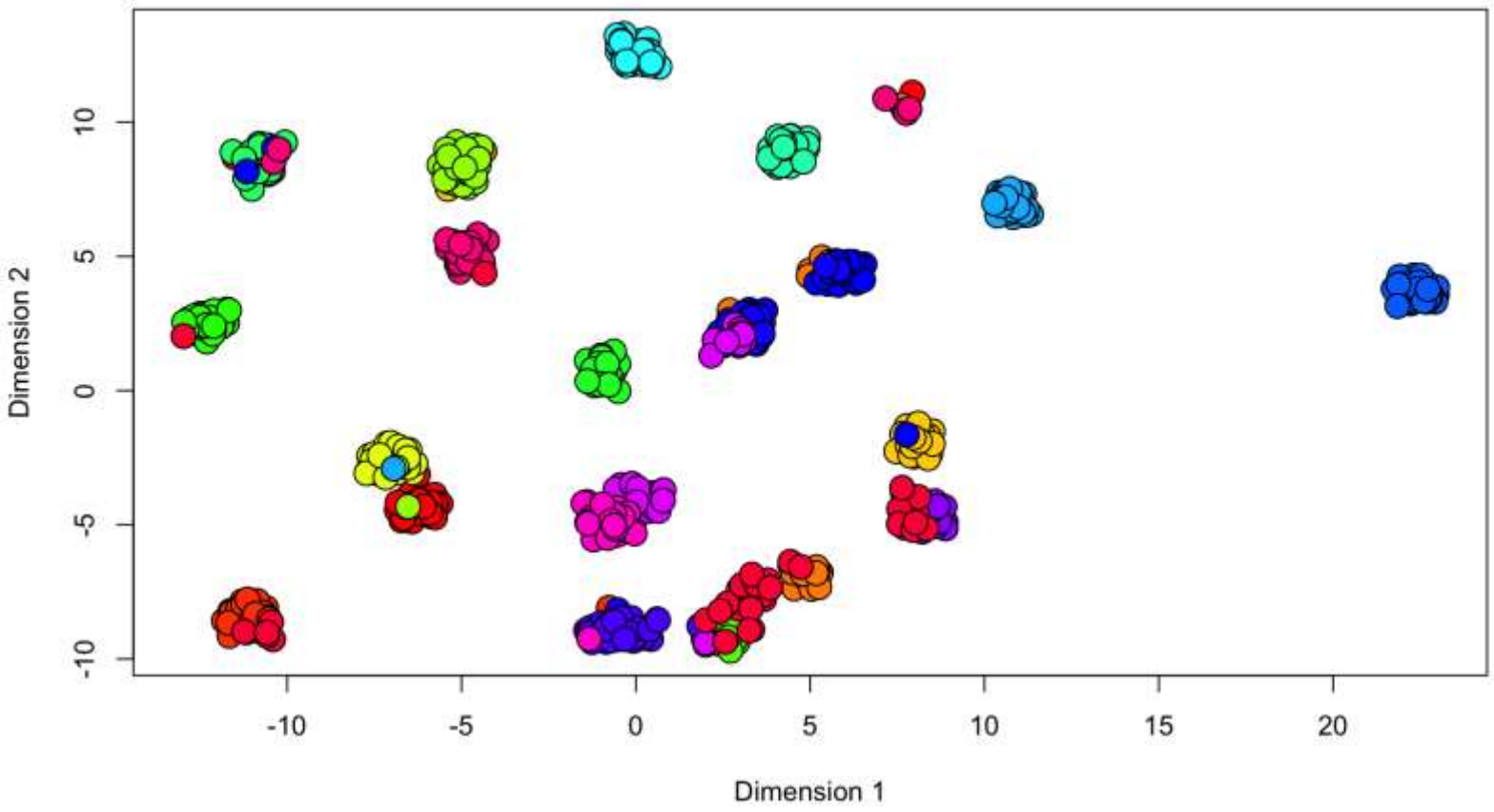
Urine from Healthy Donors

KODAMA



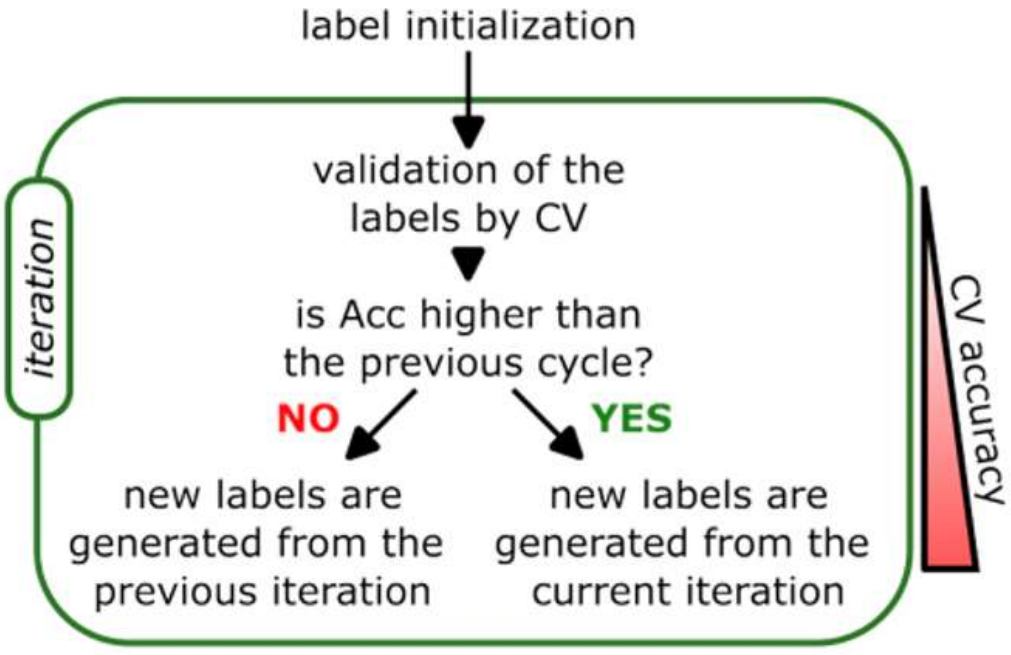
Urine from Healthy Donors

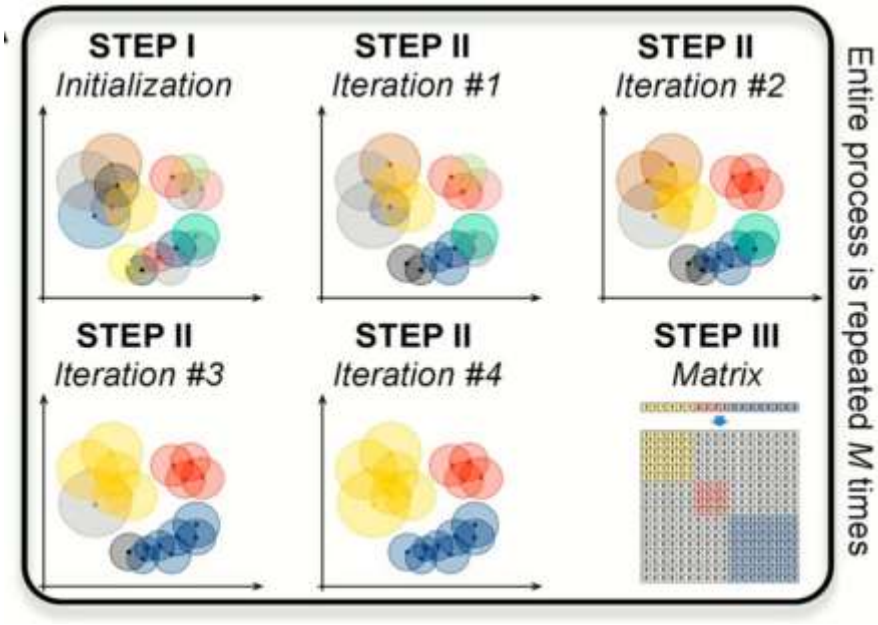
KODAMA



KODAMA

https://tkcaccia.github.io/KODAMA-Analysis/





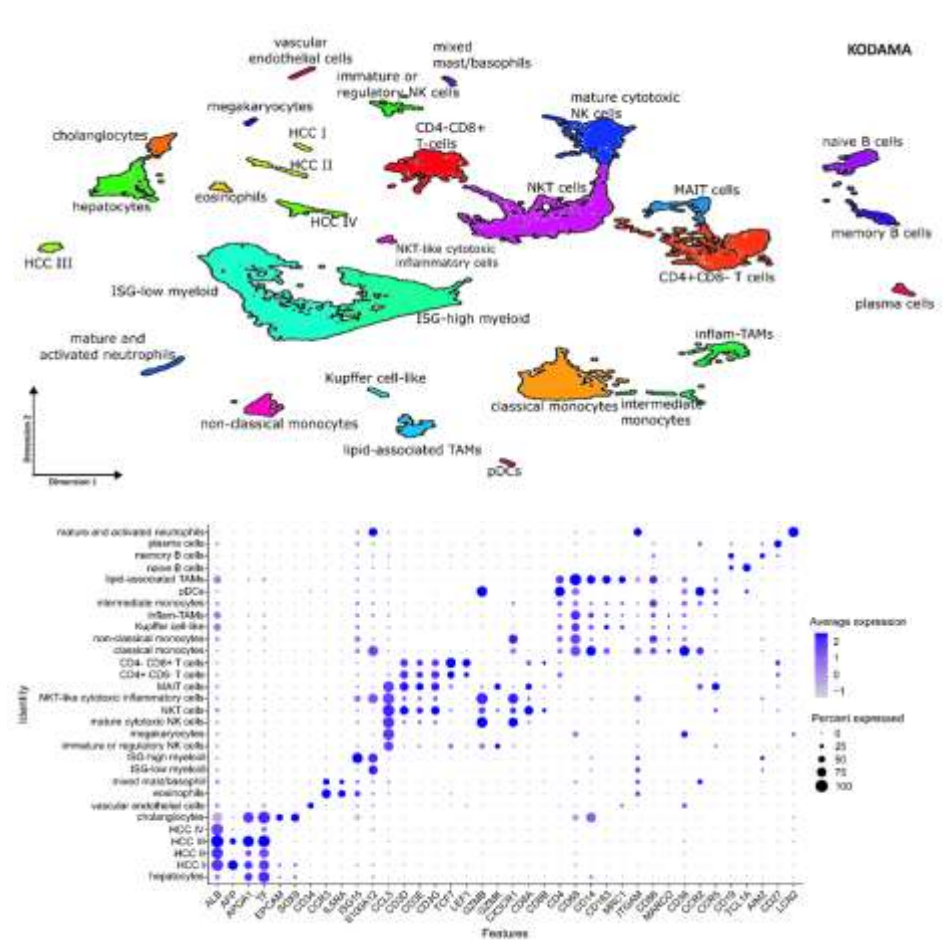
Clustering

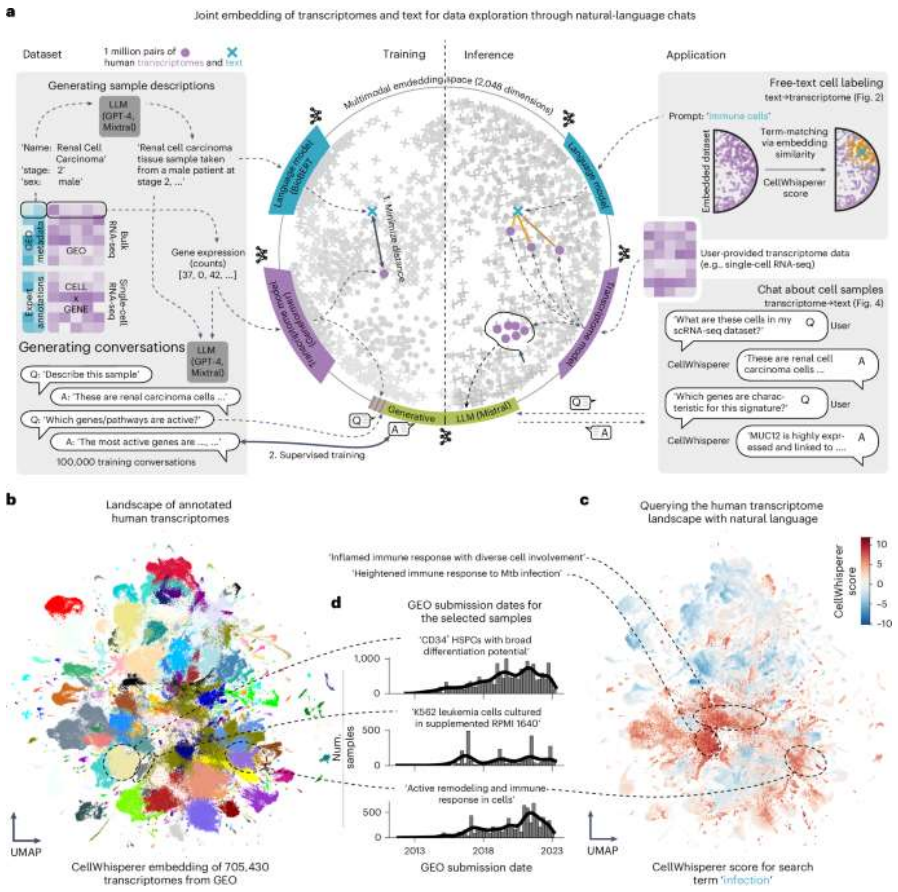
Goal: Group cells with similar transcriptional profiles into clusters representing putative cell types or states.
Approach: Build a k-nearest neighbor (kNN) graph where nodes are cells and edges connect similar expression profiles.
Algorithms:

- *Louvain:* Community detection by optimizing modularity — fast and widely used but may produce unstable partitions.
- *Leiden:* Improved version of Louvain ensuring well-connected and more robust clusters.
- *Random Walk:* Concept underlying both methods — clusters correspond to regions where random walks tend to stay longer (high graph density).

Outcome: Each cluster represents a transcriptionally coherent population, forming the basis for downstream annotation and trajectory inference.

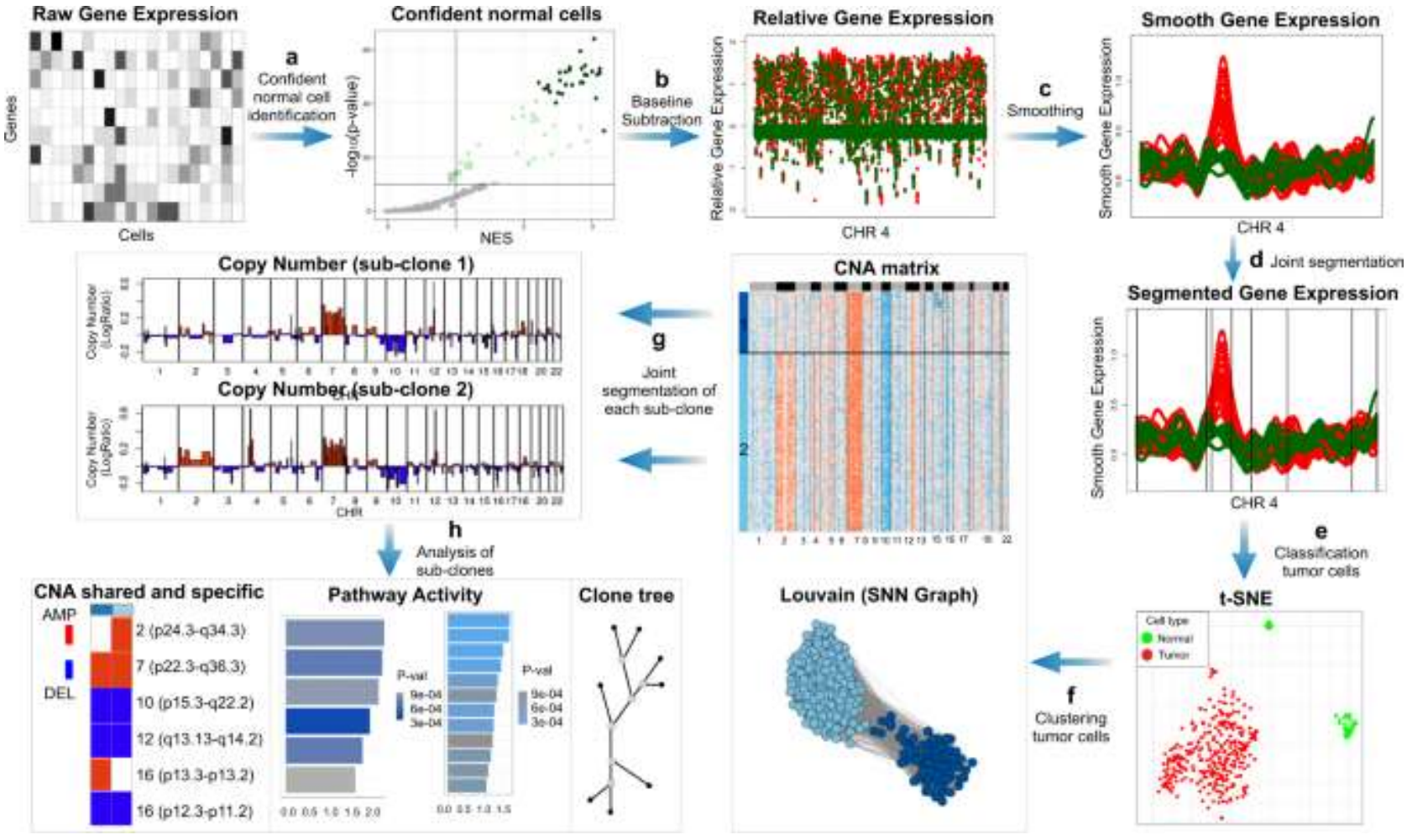
Identification



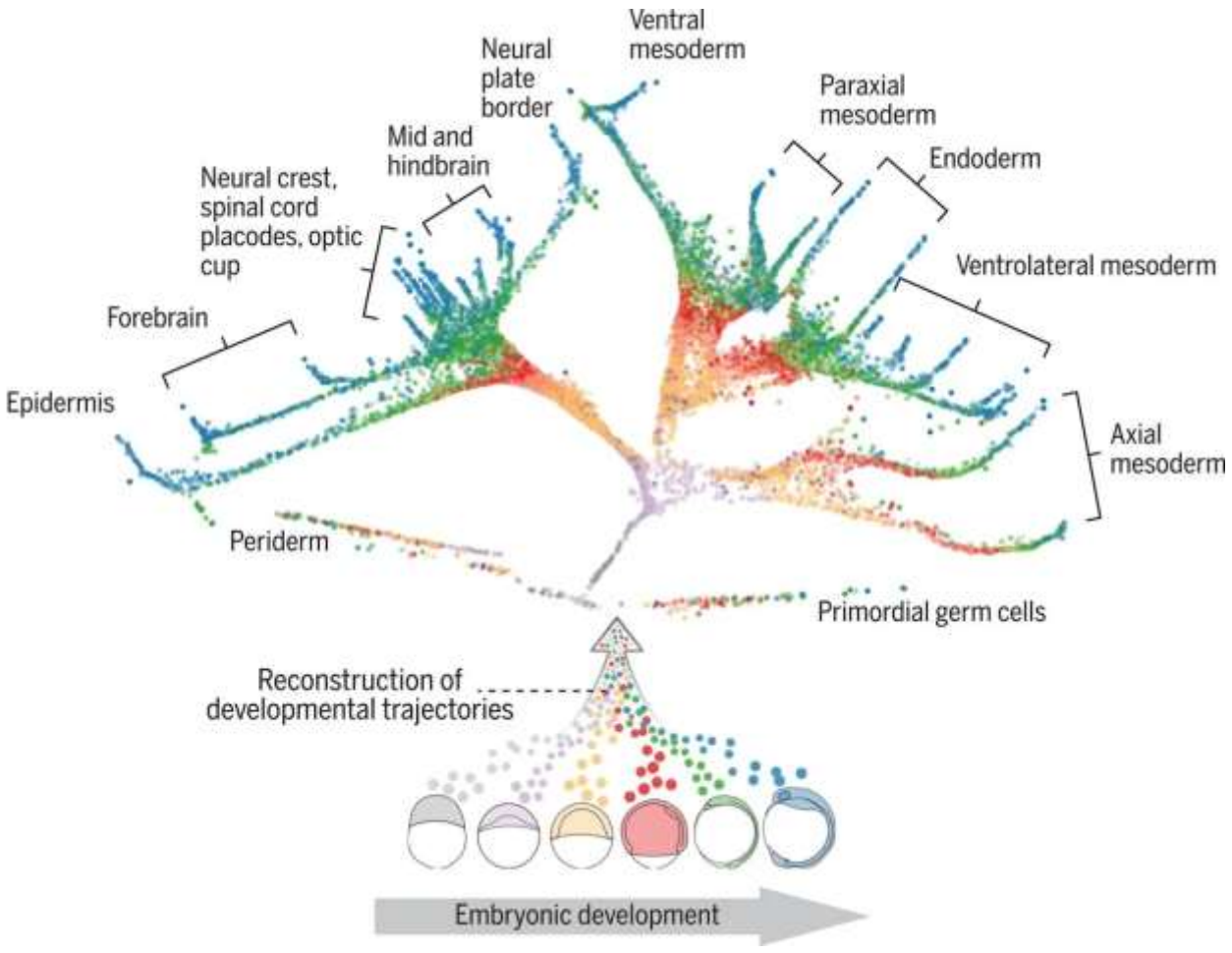


Schaefer, M., Peneder, P., Malzl, D. *et al.* Multimodal learning enables chat-based exploration of single-cell data. *Nat Biotechnol* (2025). https://doi.org/10.1038/s41587-025-02857-9

Copy number variation



Trajectory


Ventral mesoderm
Neural plate border
Paraxial mesoderm
Mid and hindbrain
Endoderm
Neural crest, spinal cord placodes, optic cup
Ventrolateral mesoderm
Forebrain
Epidermis
Axial mesoderm
Periderm
Primordial germ cells
Reconstruction of developmental trajectories
Embryonic development

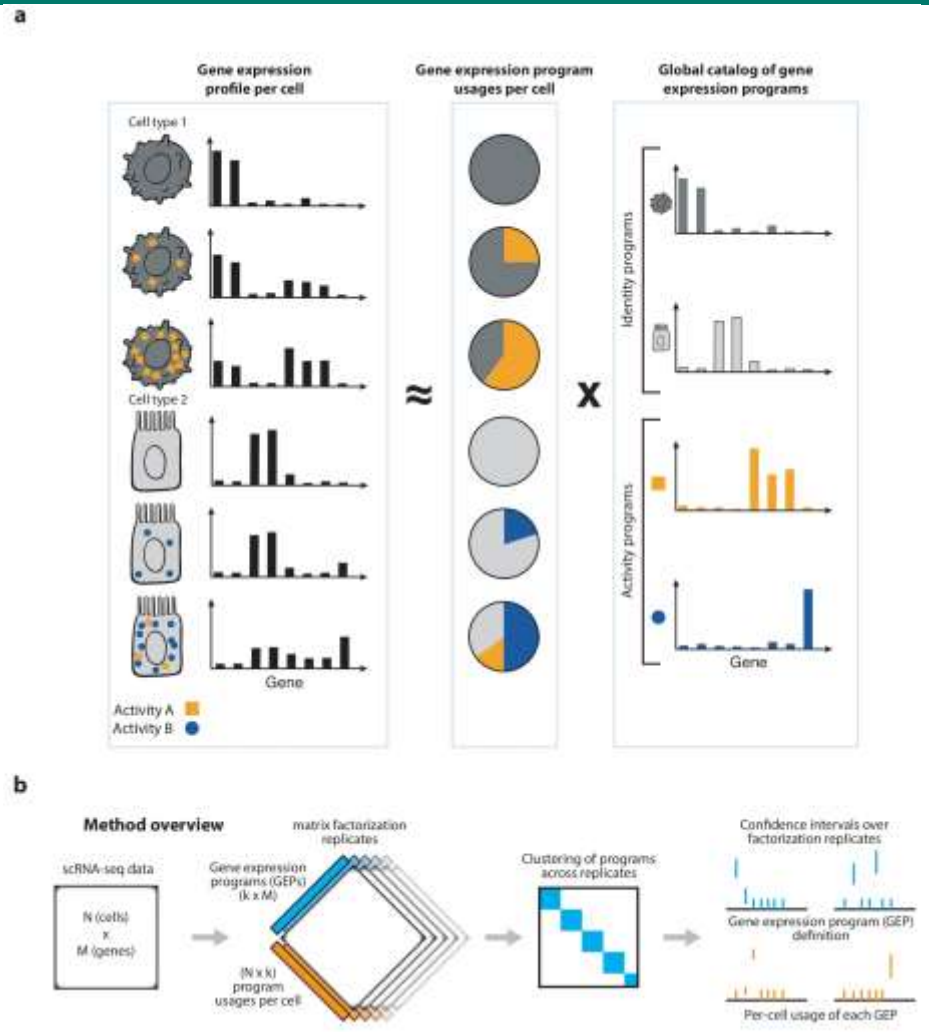
Consensus NMF for Gene Expression Programs

Goal: Identify gene expression programs (GEPs) — sets of co-expressed genes capturing distinct biological processes or cell states.

Method: Apply Non-negative Matrix Factorization (NMF) repeatedly with different random initializations or subsets of data.

Each run decomposes the expression matrix into gene weights (programs) and cell scores (program activity).

Combine results to build a consensus matrix, highlighting stable and reproducible gene–program associations.



https://github.com/dylkot/cNMF/tree/main

Special Issue

Unveiling the Complexity of Hepato-Biliary-Pancreatic Cancers: From Pathogenesis to Management

Guest Editors

Dr. Stefano Cacciatore

Dr. Luiz Zerbini

Deadline

30 June 2026

cancers

IMPACT FACTOR 4.4

Indexed in: PubMed

CITESCORE 8.8

Current group composition

Dalia Ahmed (Arturo Falaschi fellowship, Sudan)

Chiamaka Jessica Okeke (Arturo Falaschi fellowship, **Nigeria**)

Moussa Kassim (BIOTECHNET programme, **Djibouti**)

Martin Ocharo (Arturo Falaschi, **Kenya**)

Dupe Ojo (**Nigeria**).

Former members

Maria Zinga (EMPOWER fellowship, **Tanzania**)

Ebtesam A. Abdel-Shafy (Arturo Falaschi fellowship, Egypt)

Nnenna Elebo (South African National Research Foundation, **Nigeria**)

Mukhethwa Munzhedzi (ICGEB South African HDI Programme, **South Africa**)

Nancy Paola Duarte Delgado (Arturo Falaschi short fellowship, **Colombia**)

Thanks for your attention