

ICGEB TRIESTE
ITALY



ICGEB NEW DELHI
INDIA



ICGEB CAPE TOWN
SOUTH AFRICA

ICGEB

International Centre for Genetic Engineering and Biotechnology

Developing Knowledge

Multivariate analysis (part 1)

25 August 2021

www.icgeb.org

Where do we start?

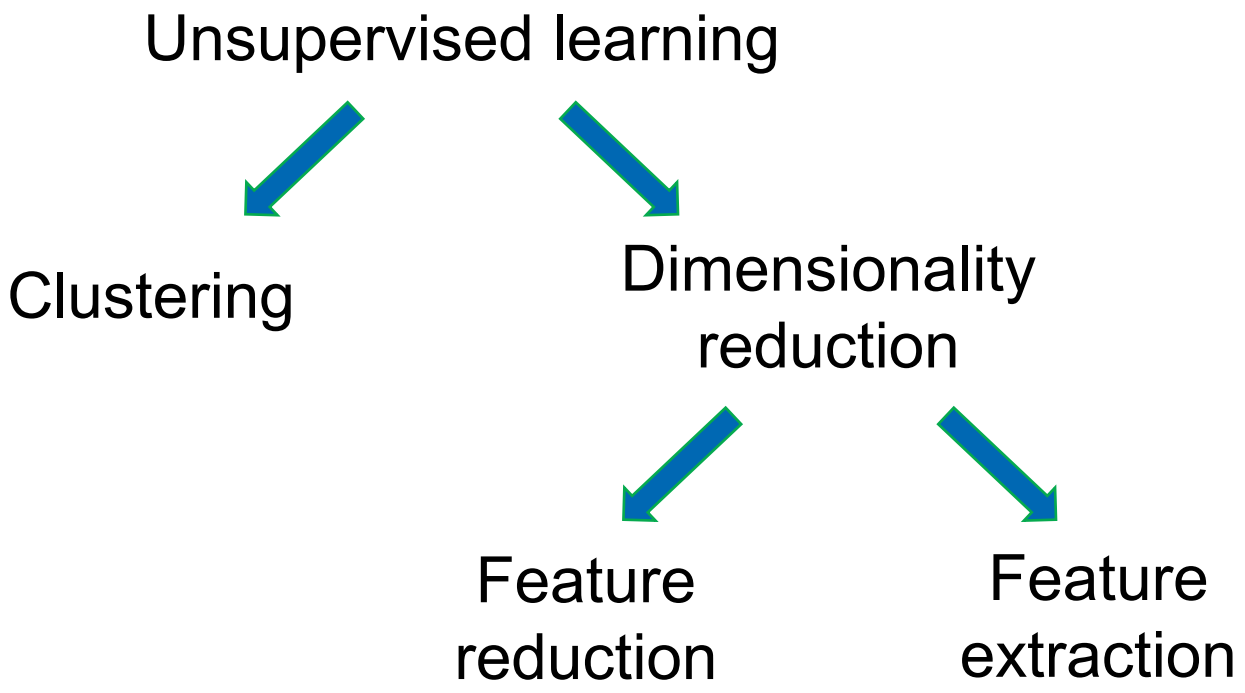
Unsupervised and supervised learning

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").

Where do we start?

Unsupervised learning



Unsupervised learning

Clustering

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.

Unsupervised learning

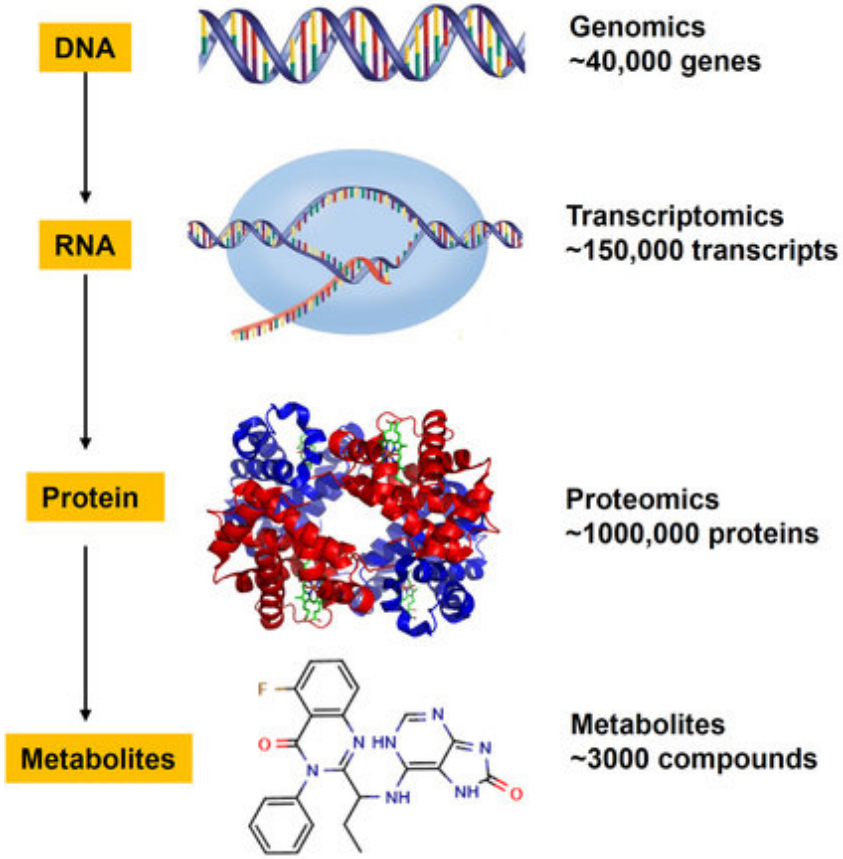
Dimensionality reduction

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.

Metabolomics

Definition

In analogy to genomics and proteomics, the term metabolomics can be defined as the study of the complete ensemble of all small molecules (molecular weight (MW) < 1500 Da) formed by numerous biosynthetic and catabolic pathways within a biological system or originating from host-specific microbes and the intake of food nutrients and pharmaceuticals, which are present in a cell, tissue, or biofluids such as urine, blood, or saliva, in the context of a physiological or pathological condition.



Metabolomics

History

This urine wheel was published in 1506 by Ullrich Pinder, in his book Epiphanie Medicorum. The wheel describes the possible colors, smells and tastes of urine, and uses them to diagnose disease.

Ibn Sina was doing metabolomics about 1000 years ago on urine:

- Color;
- Concentration;
- Clarity;
- Sediment;
- Volume;
- Odor;
- Foam or Froth.

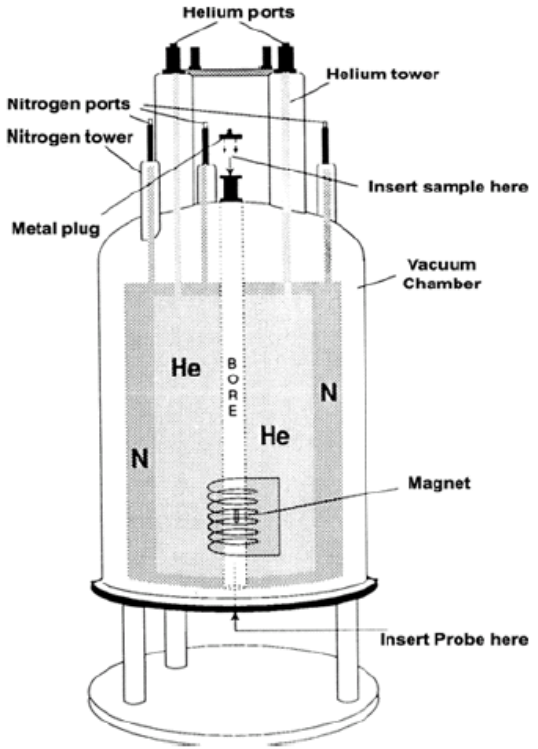


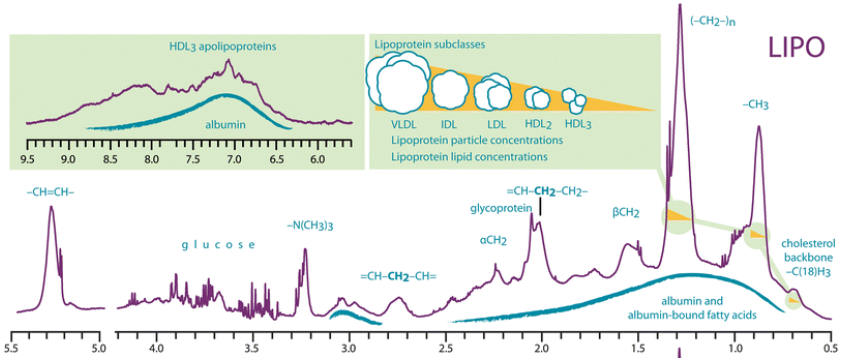
Nicholson, J. K. & Lindon, J. C. Nature 455, 1054–1056

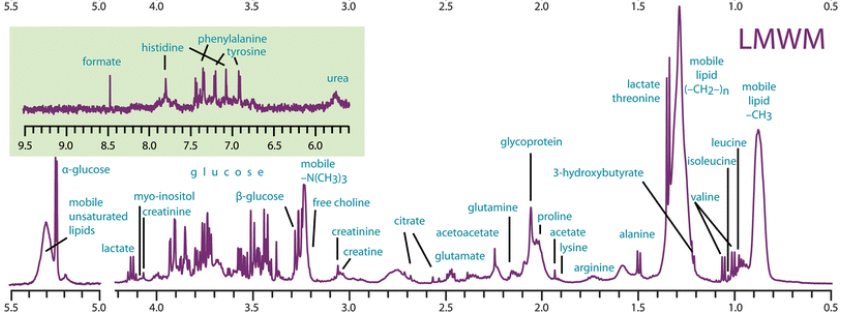
Metabolic Profiling Methods

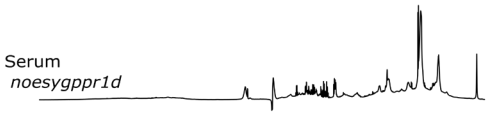
Nuclear Magnetic Resonance (NMR) spectroscopy

NMR uses radio-frequency (RF) radiation and magnetic fields. RF radiation is used to stimulate nuclei present within molecules. The information obtained is displayed as a spectrum. The horizontal axis is the chemical shift (delta, in units of ppm), which is a measure of the position at which RF absorption occurs relative to an internal standard (tetramethylsilane, TMS). The vertical axis is the intensity of the absorption. As with other spectral techniques, compounds have characteristic spectra. More than 100 metabolites occur in plants at levels high enough for analysis by NMR, so NMR spectra of mixtures contain many peaks.

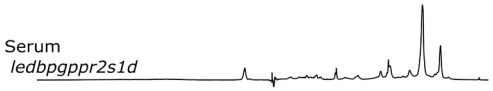




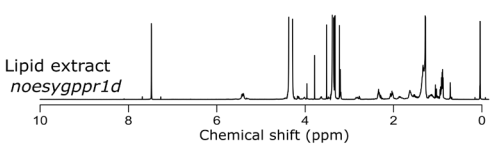




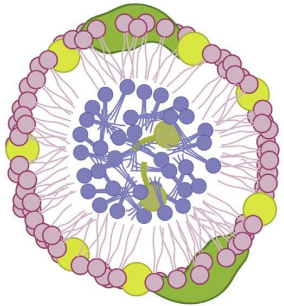
Ketone bodies and carboxylic acids
(acetate, acetoacetate, 3-hydroxybutyrate, 2-hydroxybutyrate, formate)
Energetic related metabolites
(glucose, lactate, pyruvate, citrate)
Inflammation markers
(GlycA, GlycB, albumin)
Amino acids
(alanine, glutamine, glycine, histidine, isoleucine, leucine, valine, phenylalanine, tyrosine, glutamate, threonine)
Fluid balance
(creatinine, creatine)
Vitamins
(ascorbate)
Others
(mannose, ethanol, lipids)

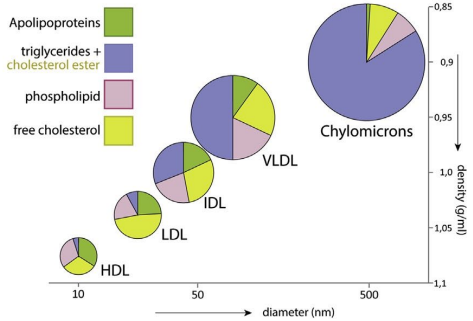


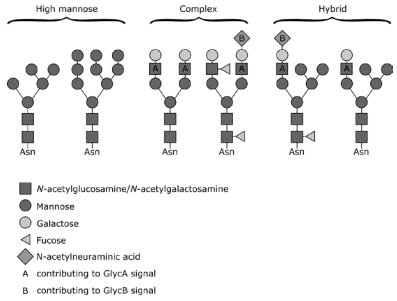
Lipoprotein subclasses (VLDL, IDL, LDL, HDL)
- particle number
- dimension
- lipid composition (triglycerides, cholesterol)

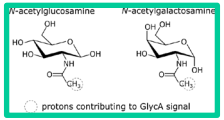


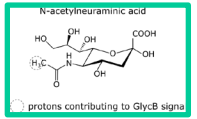
Fatty acids
(saturated fatty acid, omega-9 fatty acid, omega-7 fatty acid, omega-6 fatty acid, omega-3 fatty acid, polyunsaturated fatty acid, monounsaturated fatty acid, linoleic acid)
Triglycerides
Phospholipids
(Sphingomyelin, Phosphatidylcholine, Glycerophospholipid)
Conjugated and free sterols
(free cholesterol, esterified cholesterol)





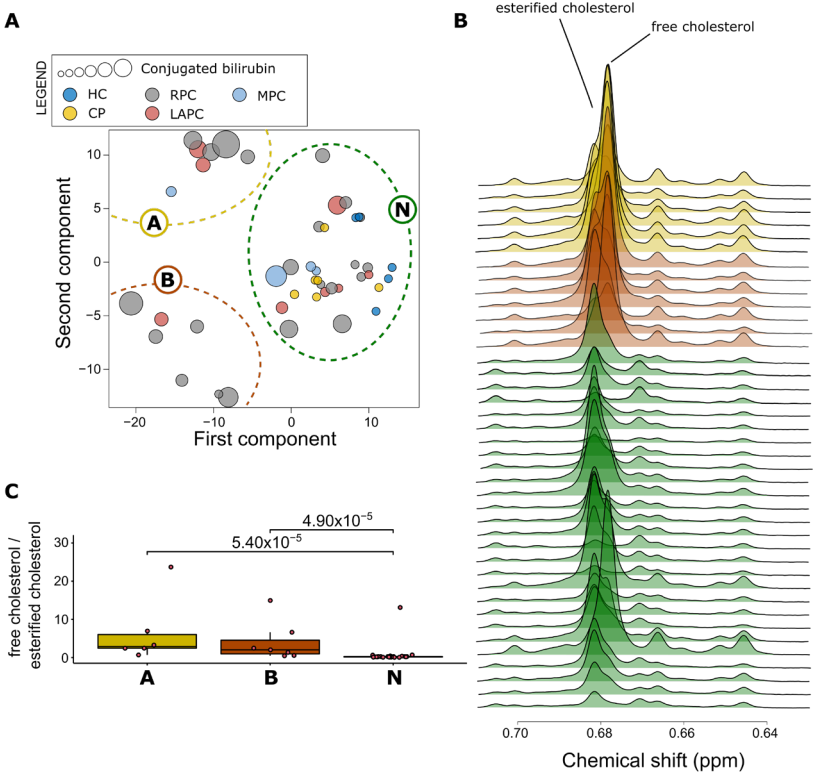


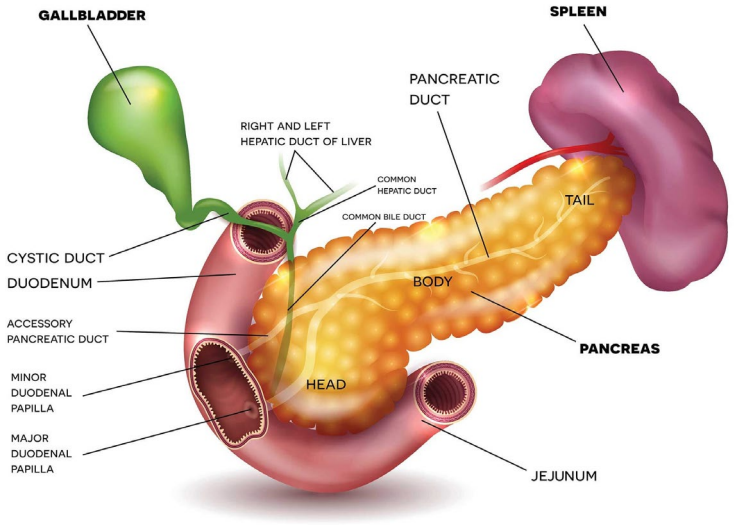




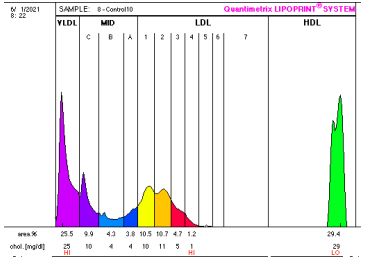
Pancreatic cancer

Obstructive Jaundice

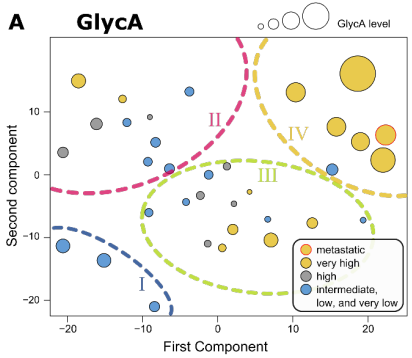


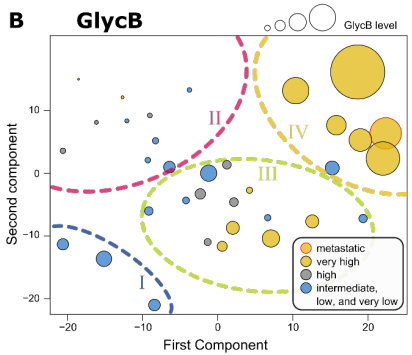


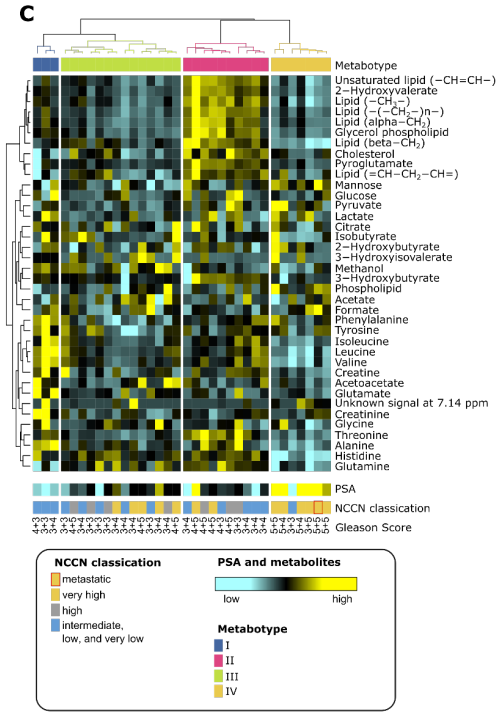


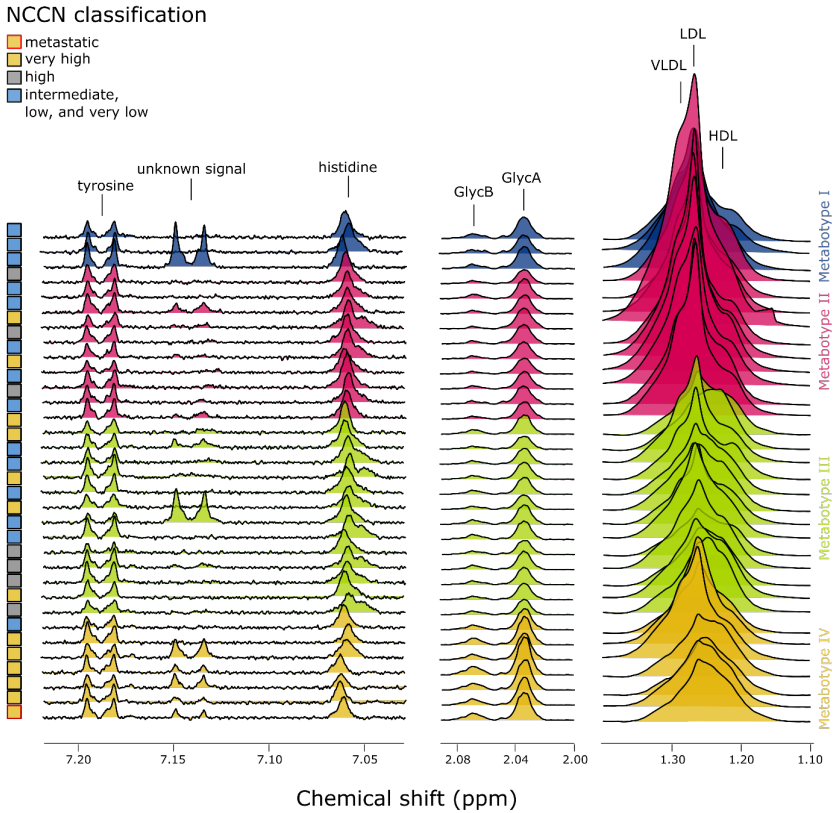


Inflammatory metabolic profile of South African patients with prostate cancer







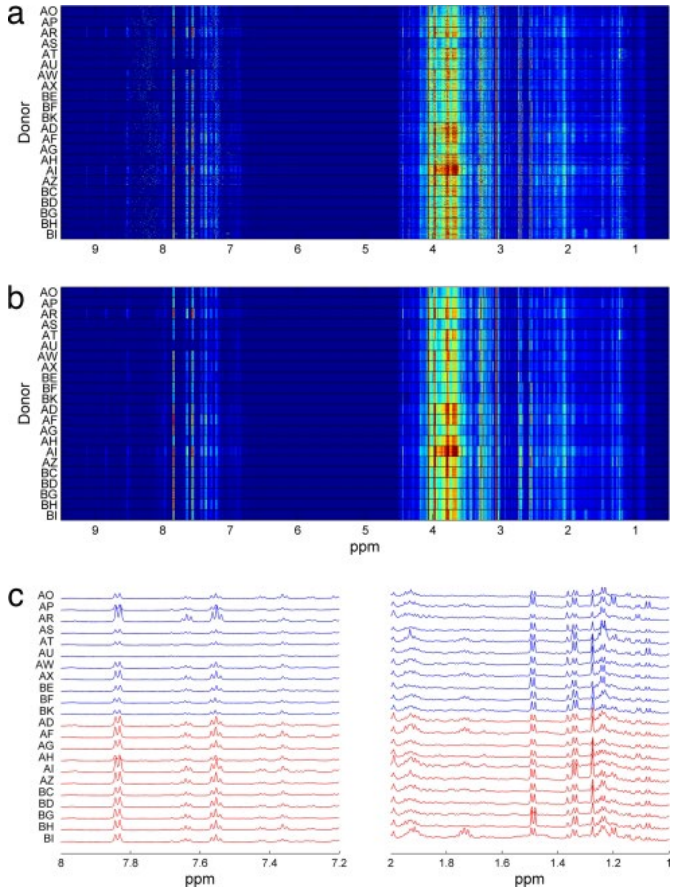


Cacciatore S, Wium M, Licari C, Ajayi-Smith A, Masieri L, Anderson C, Salukazana AS, Kaestner L, Carini M, Carbone GM, Catapano CV, Loda M, Libermann TA, Zerbini LF. Inflammatory metabolic profile of South African patients with prostate cancer. Cancer Metab . 2021 Aug 3;9(1):29.

Case study

Metabolomic data

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.



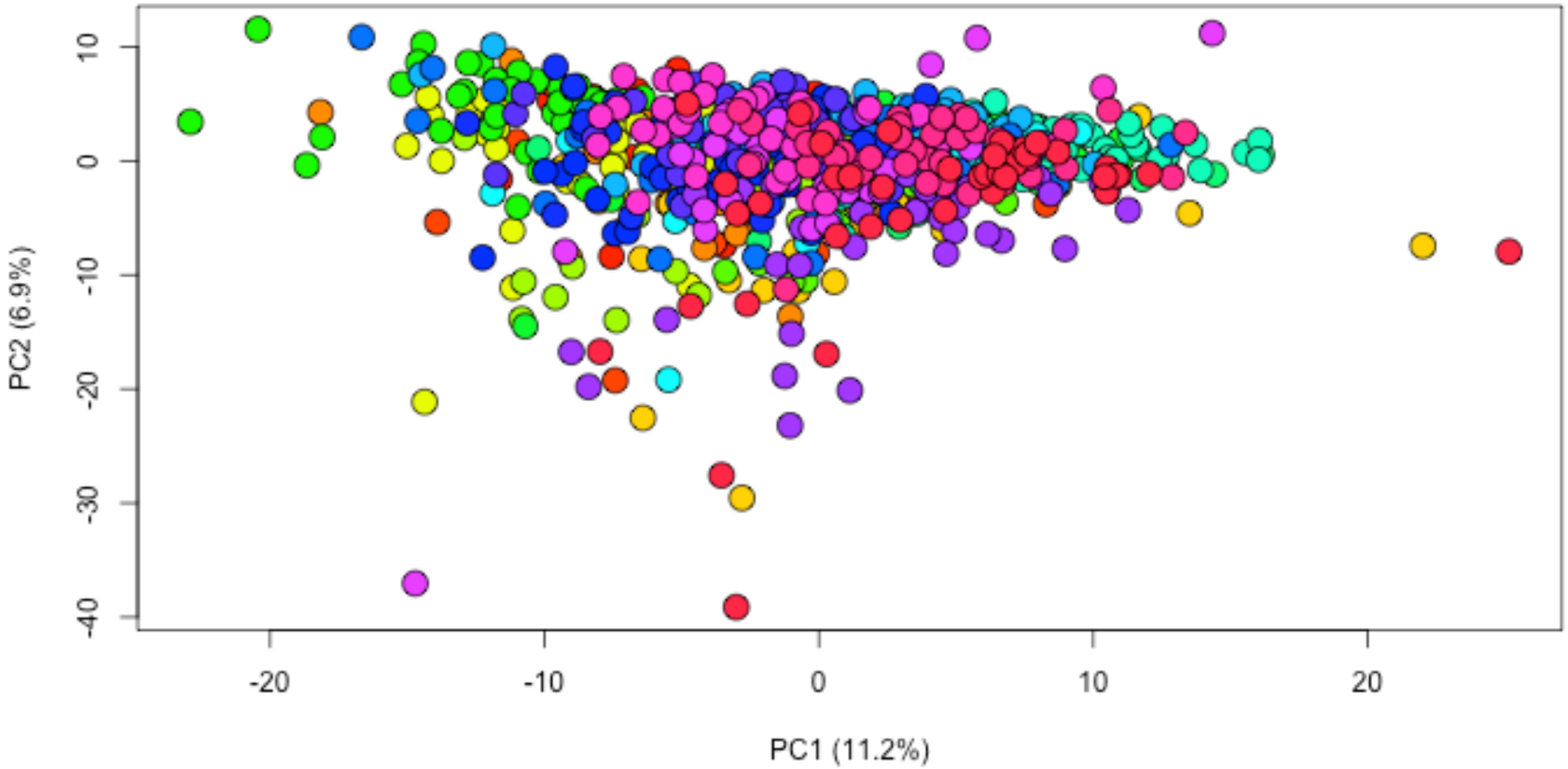
Proc Natl Acad Sci U S A. 2008 Feb 5;105(5):1420-4.

Evidence of different metabolic phenotypes in humans

Michael Assfalg, Ivano Bertini, Donato Colangiuli, Claudio Luchinat, Hartmut Schäfer, Birk Schütz, Manfred Spraul

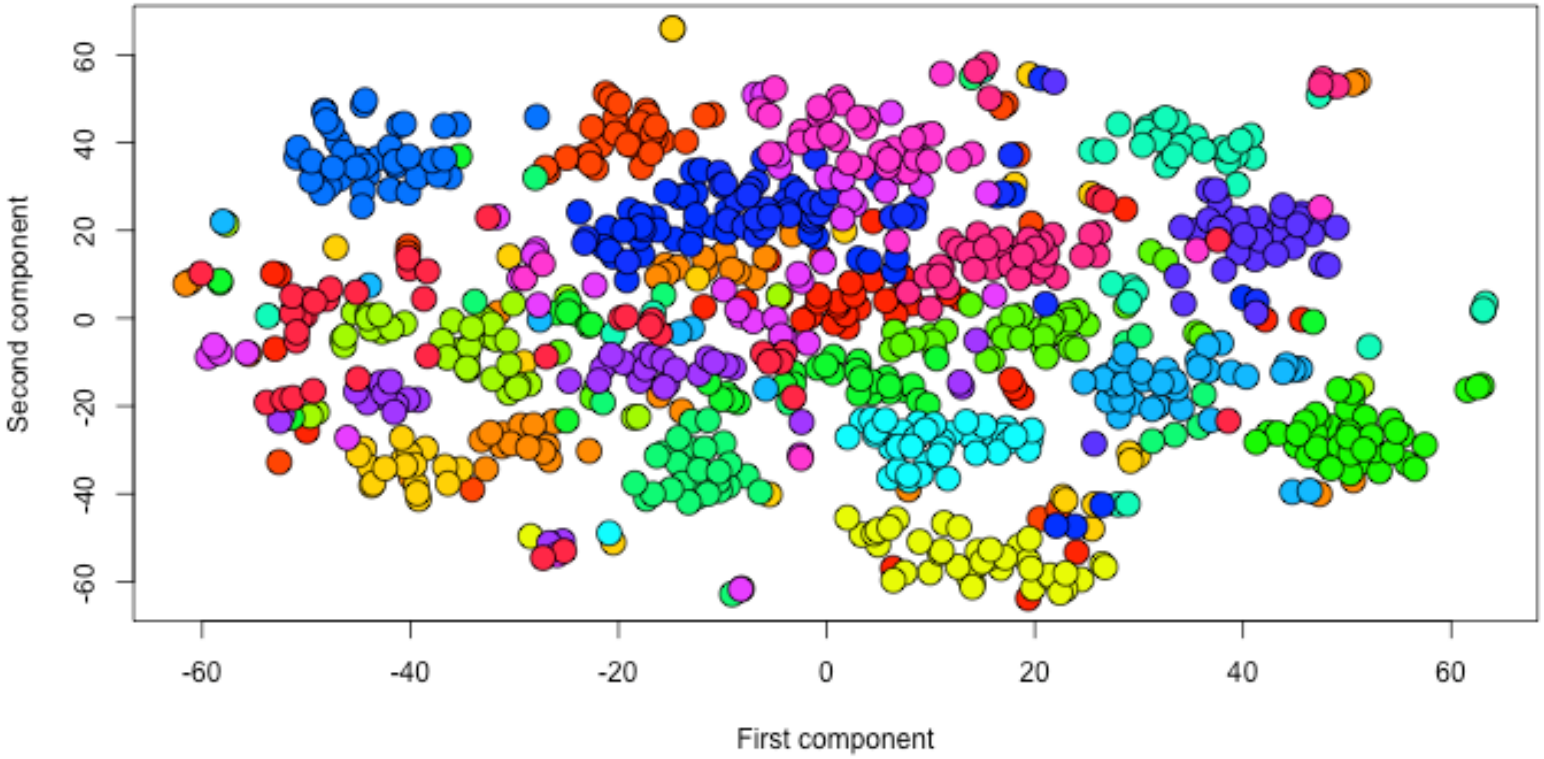
METref

Principal Component Analysis



MEtRef

t-SNE


Second component
60
40
20
0
-20
-40
-60
-60
-40
-20
0
20
40
60
First component

Knowledge discovery by accuracy maximization

Stefano Cacciatore[a,b,c], Claudio Luchinat[a,d,1], and Leonardo Tenori[d]

[a]Magnetic Resonance Center (CERM) and Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Italy; [b]Department of Medical Oncology, Dana–Farber Cancer Institute, Harvard Medical School, Boston, MA 02115; [c]Metabolomics Platform, Spanish Biomedical Research Center in Diabetes and Associated Metabolic Disorders, Rovira i Virgili University, 43007 Tarragona, Spain; and [d]FiorGen Foundation, 50019 Sesto Fiorentino, Italy

Bioinformatics, 33(4), 2017, 621–623
doi: 10.1093/bioinformatics/btw705
Advance Access Publication Date: 30 November 2016
Applications Note

OXFORD

Freely available as R package

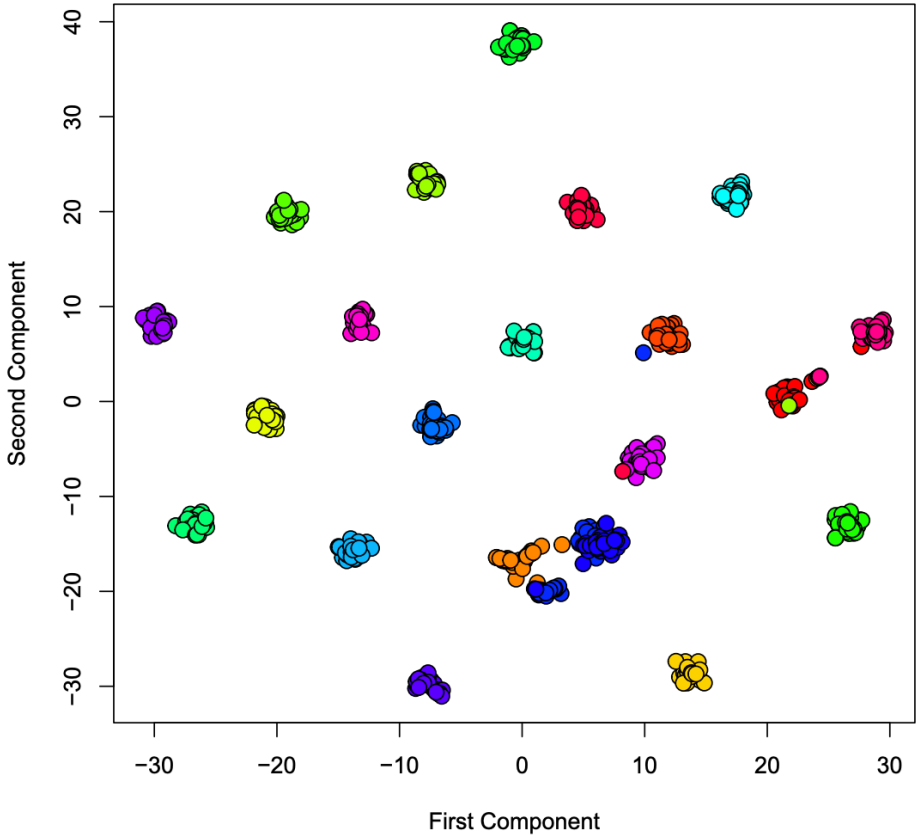
Data and text mining

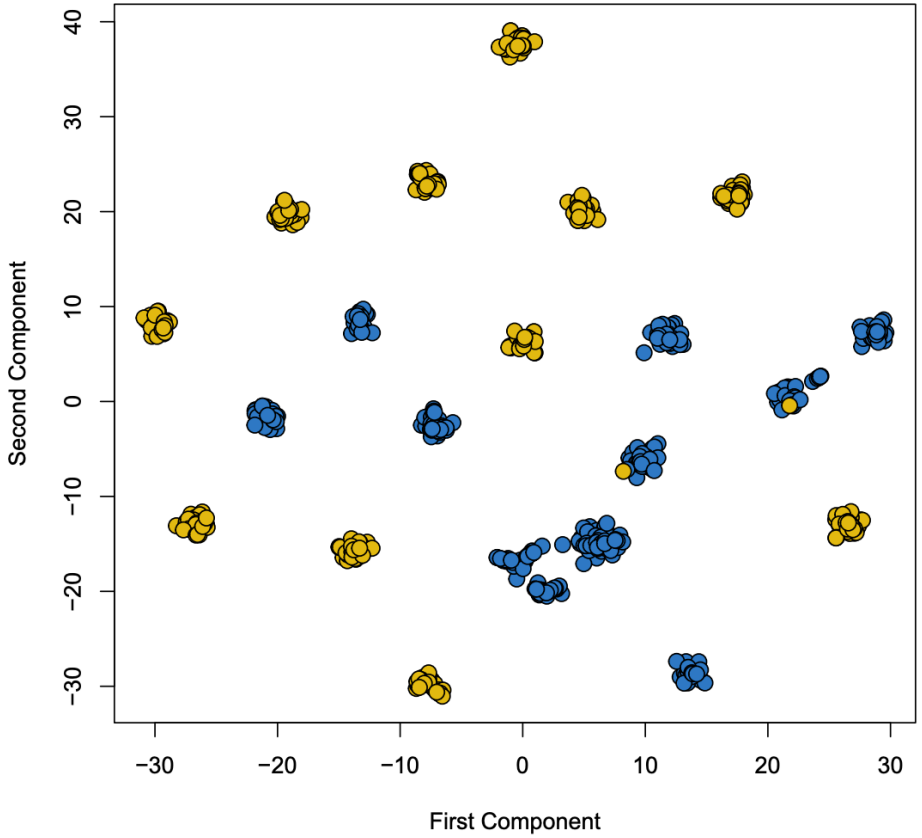
KODAMA: an R package for knowledge discovery and data mining

Stefano Cacciatore[1,*], Leonardo Tenori[2], Claudio Luchinat[3], Phillip R. Bennett[1] and David A. MacIntyre[1]

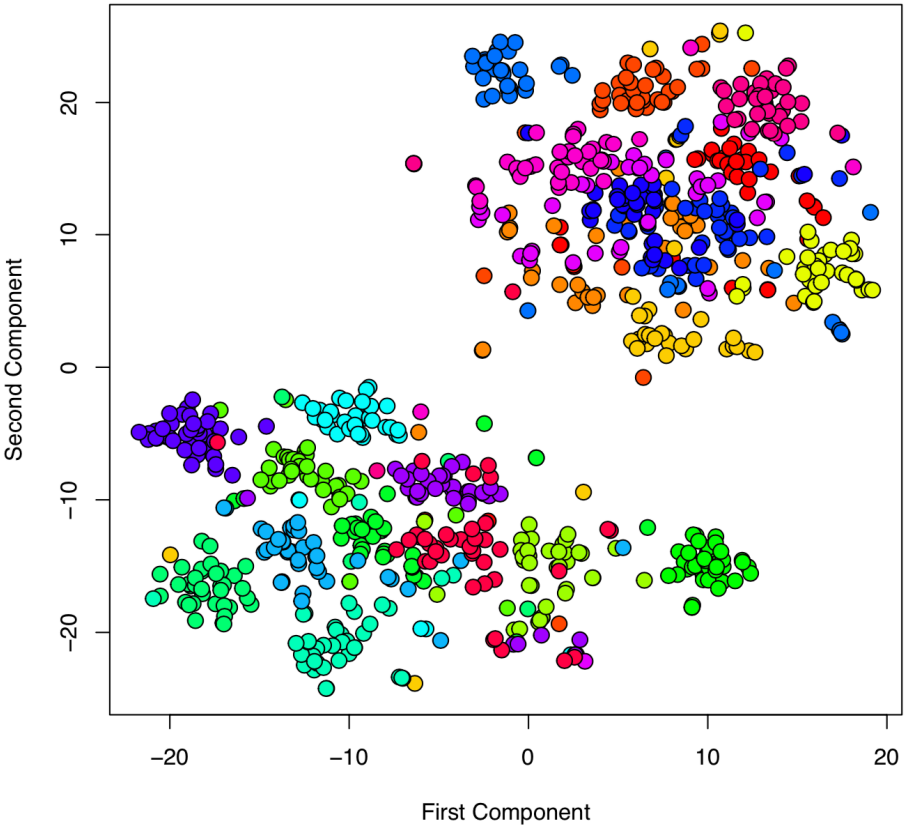
[1]Institute of Reproductive and Developmental Biology, Imperial College London, London, UK, [2]Department of Clinical and Experimental Medicine and [3]Centro Risonanze Magnetiche, University of Florence, Florence, Italy

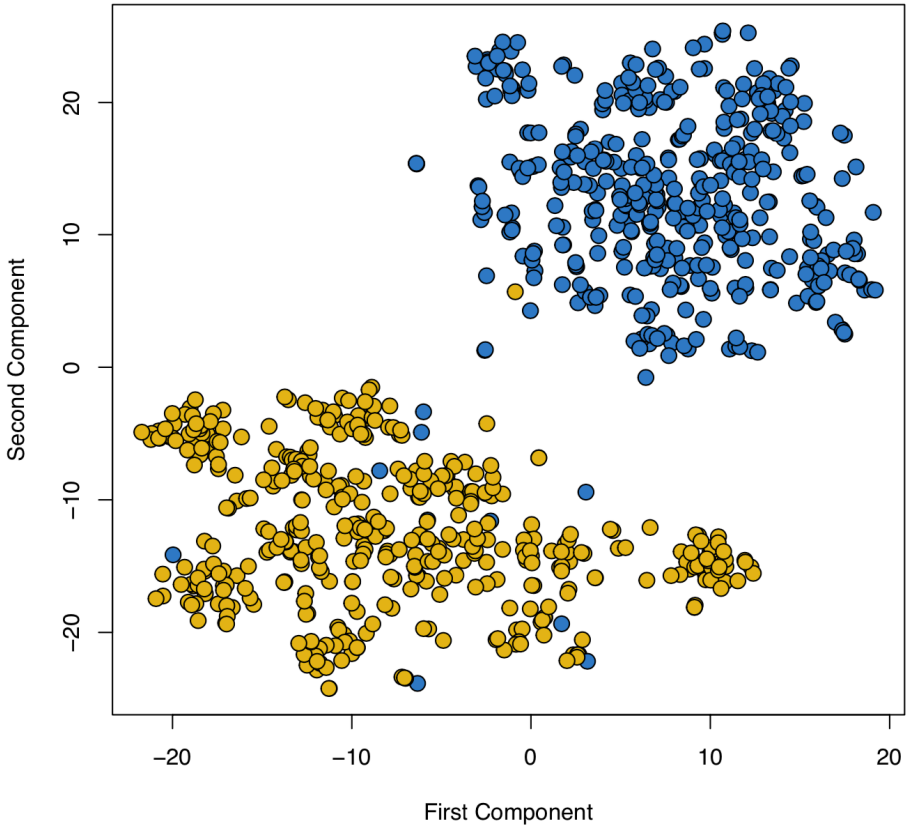
MEtRef
KODAMA





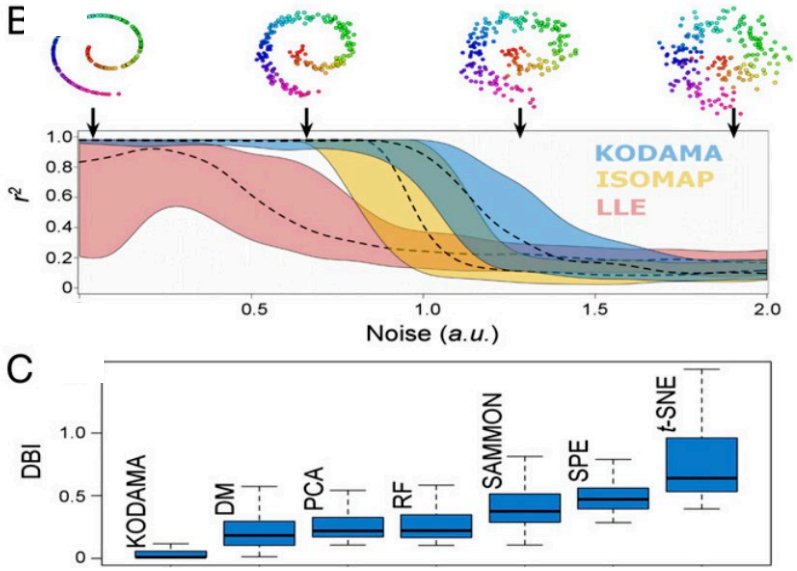
MEtRef
KODAMA

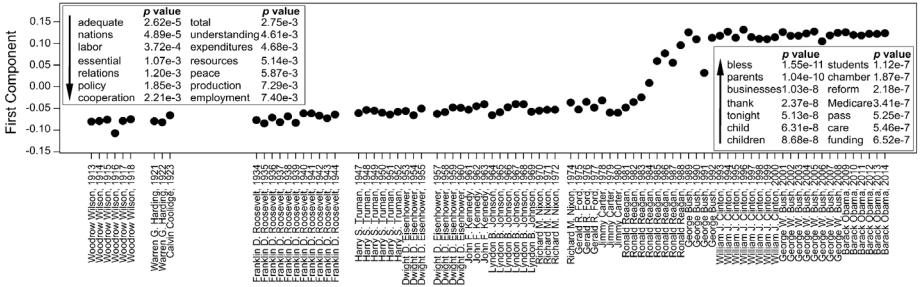




KODAMA

Knowledge discovery by accuracy maximization





Feature extraction methods:

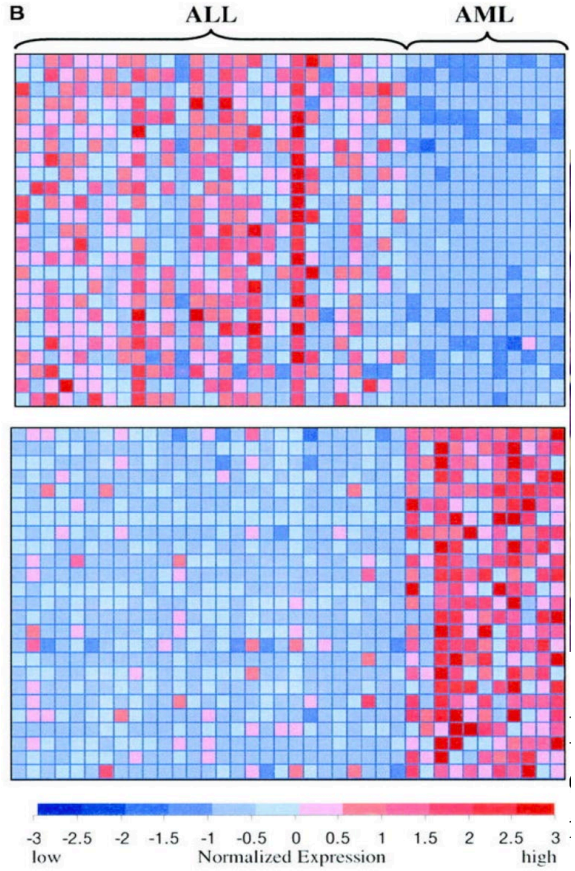
- **ISOMAP** (*Science 2000*)
- **LLE** (*Science 2000*)
- **Random Forest** (*Mach Learn 2001*)
- **t-SNE** (*J Mach Learn Res 2008*)

Clustering methods:

- **Affinity Propagation** (*Science 2007*)

Machine learning

Beginning and future directions



Science

Current Issue First release papers Archive About Submit manuscript

REPORTS

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

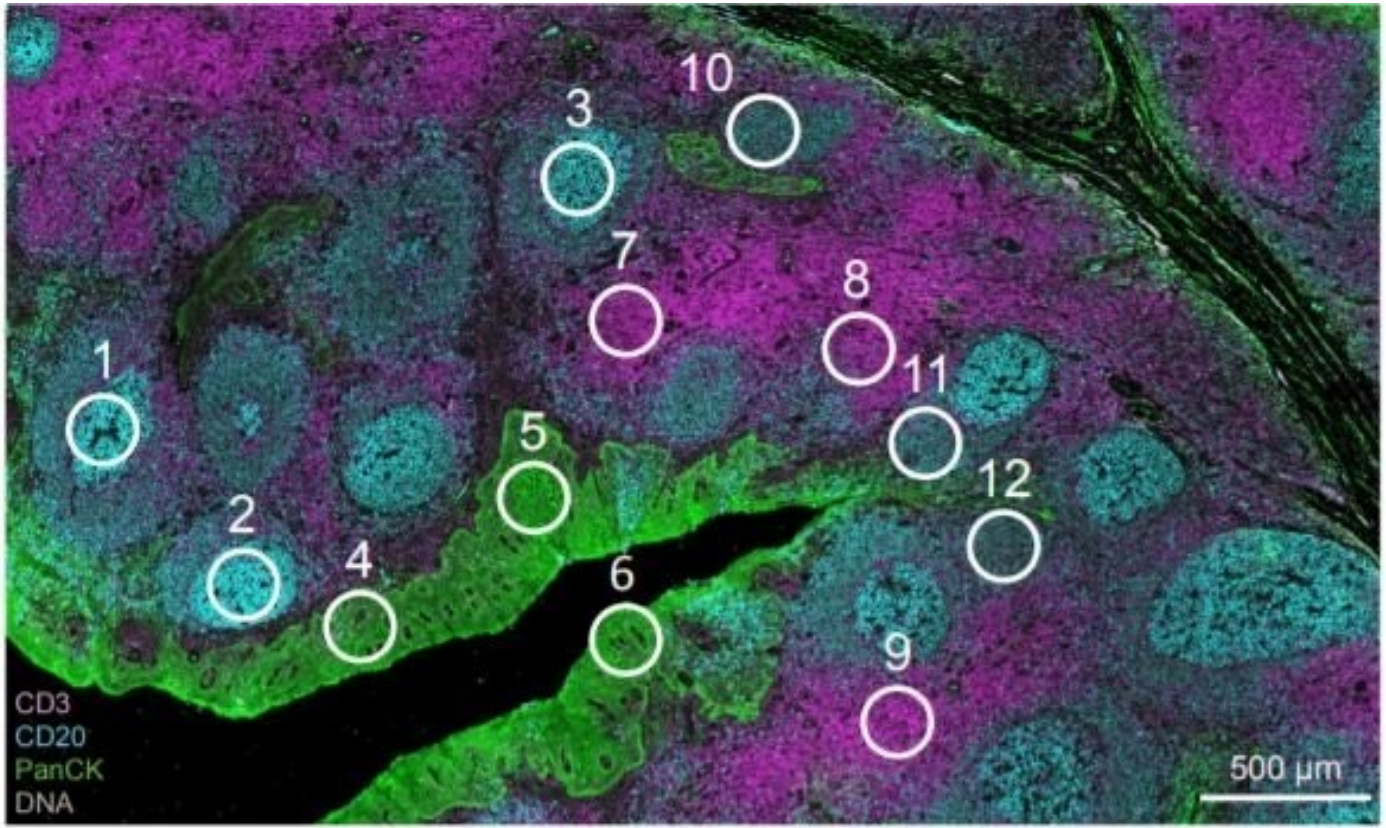
T. R. GOLUB, D. K. SLONIM, P. TAMAYO, C. HUARD, M. GAASENBEEK, J. P. MESIROV, H. COLLER, M. L. LOH, J. R. DOWNING, [...] E. S. LANDER +3 authors Authors Info & Affiliations

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.

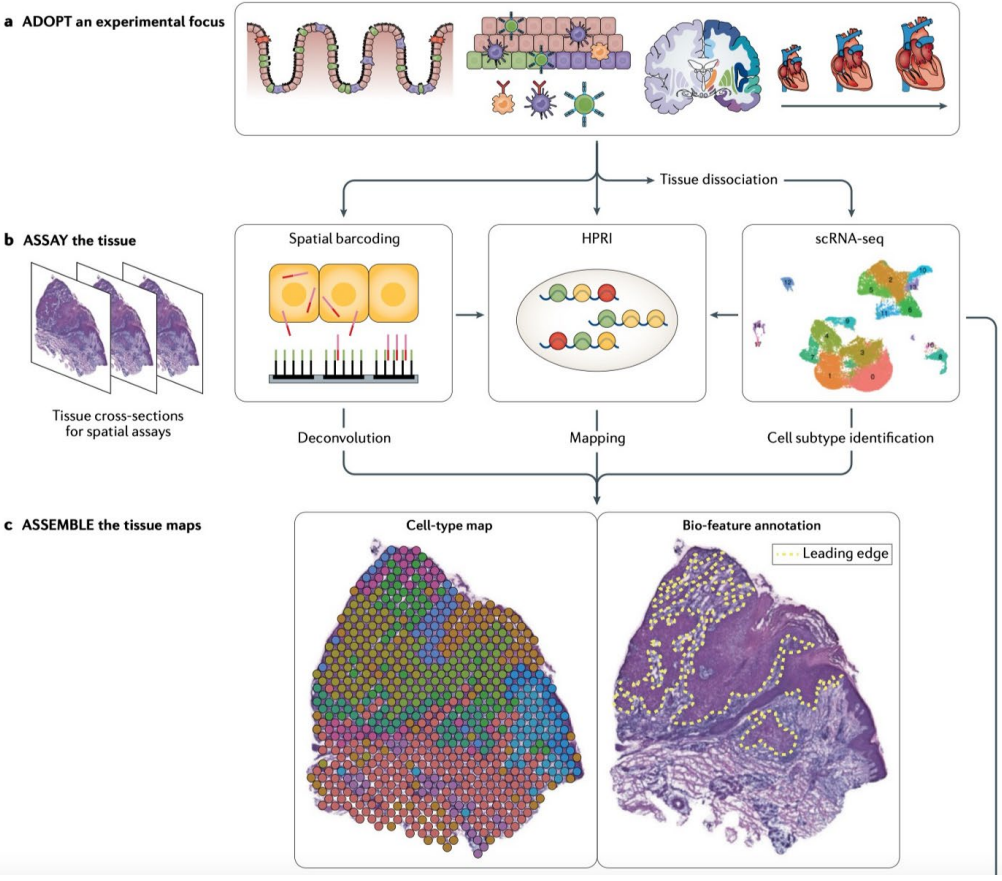
Machine learning

Beginning and future directions



Machine learning

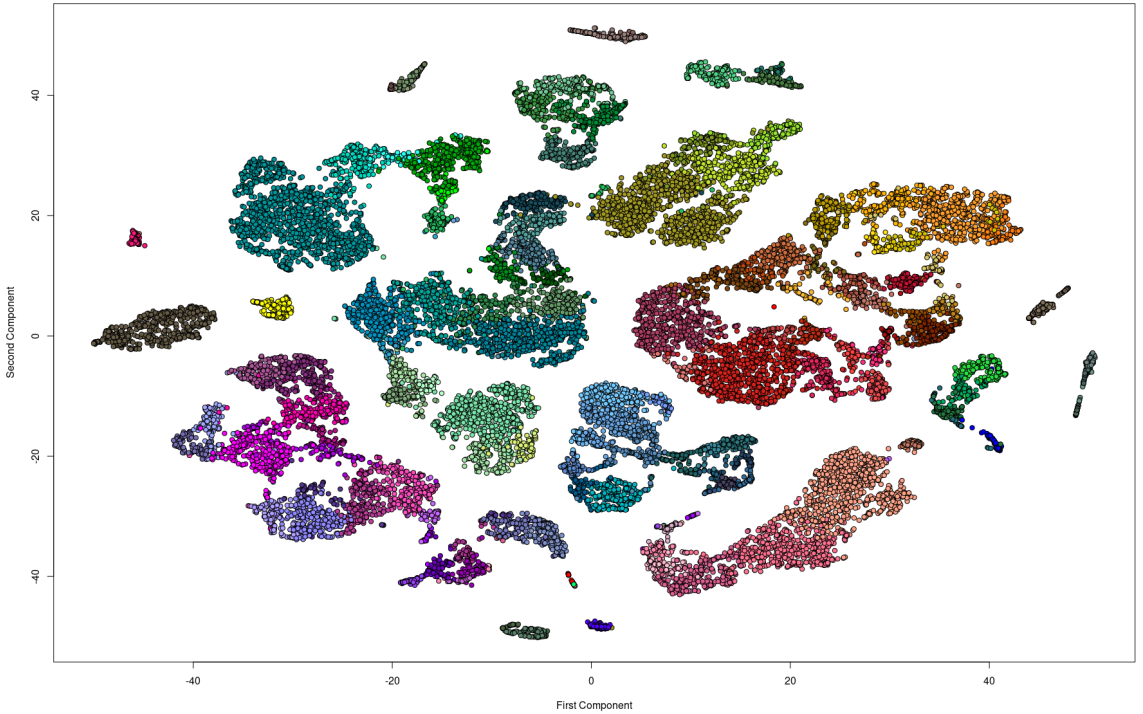
Beginning and future directions



Single-cell transcriptomic

Knowledge discovery by accuracy maximization

The data set from Tasic et al. encompasses 23,822 cells from adult mouse cortex, split by the authors into 133 clusters with strong hierarchical organisation. A standard preprocessing pipeline consisting of sequencing depth normalisation, feature selection, log-transformation, and reducing the dimensionality to 50 PCs was applied as described by Kobak & Berens in The art of using t-SNE for single-cell transcriptomics.





ICGEB TRIESTE
ITALY



ICGEB NEW DELHI
INDIA



ICGEB CAPE TOWN
SOUTH AFRICA

ICGEB

International Centre for Genetic Engineering and Biotechnology

Developing Knowledge

Multivariate analysis (part 2)

26 August 2021

www.icgeb.org

Where do we start?

Unsupervised and supervised learning

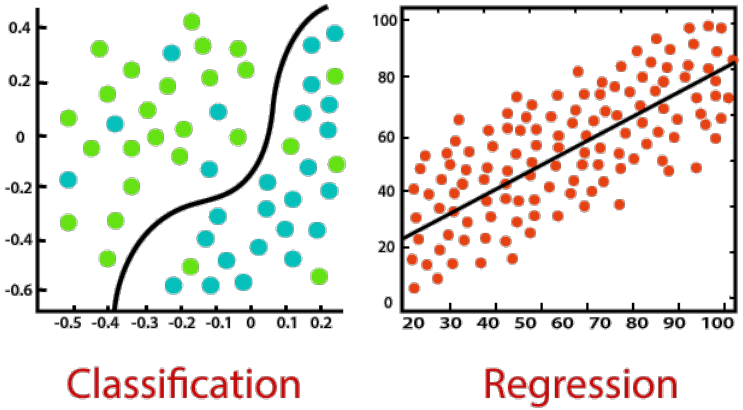
Supervised learning is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over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

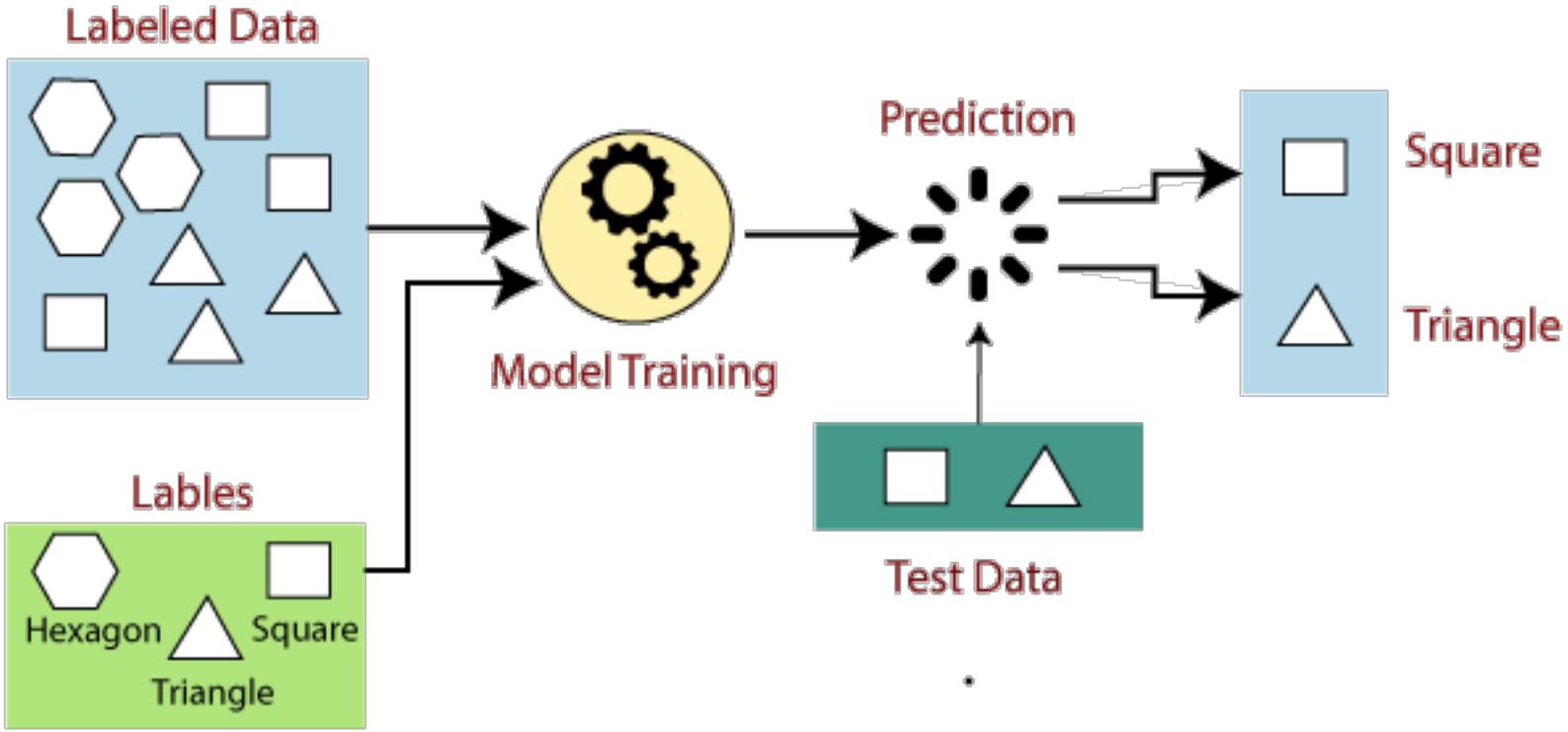
Classification and regression

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.



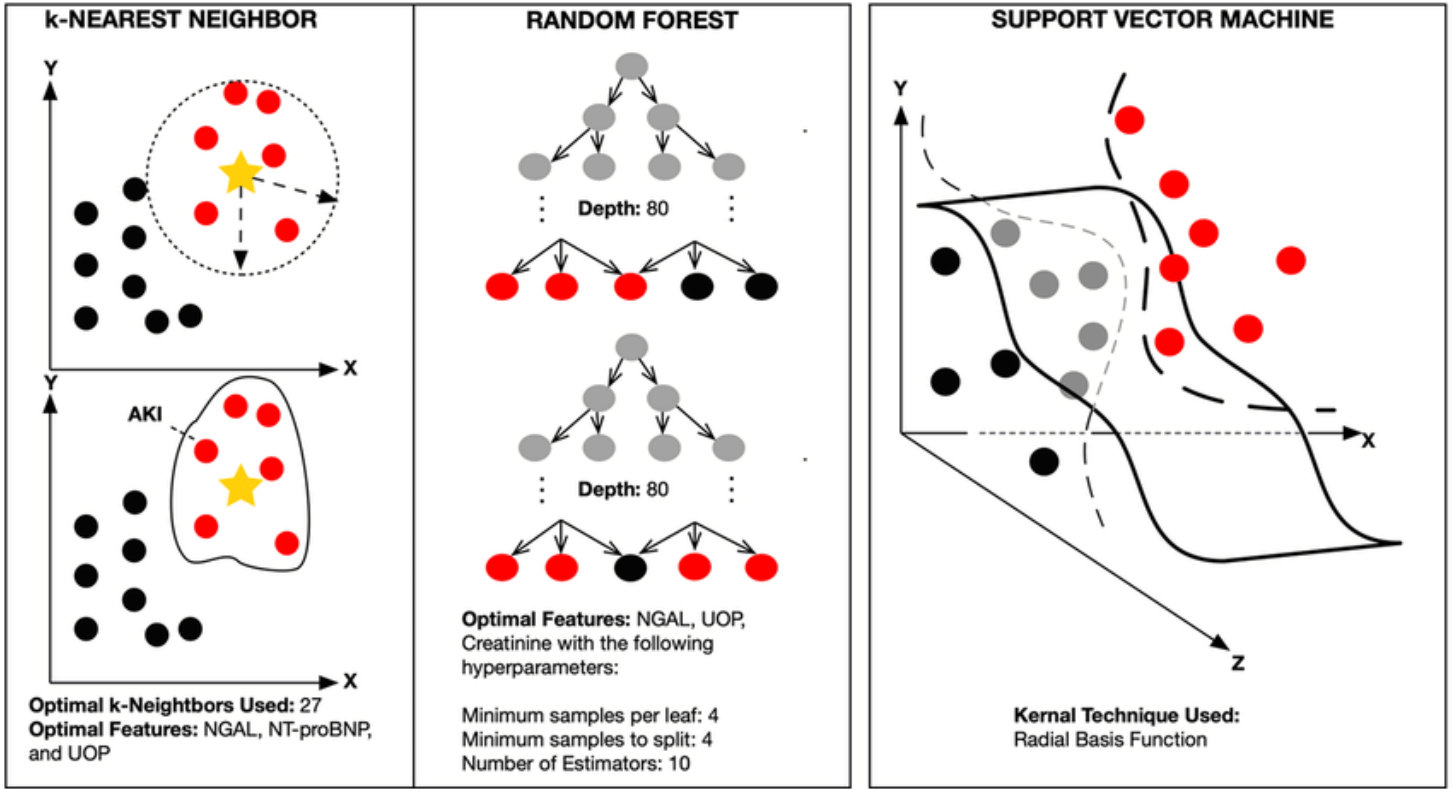
Supervised learning

Classification and regression



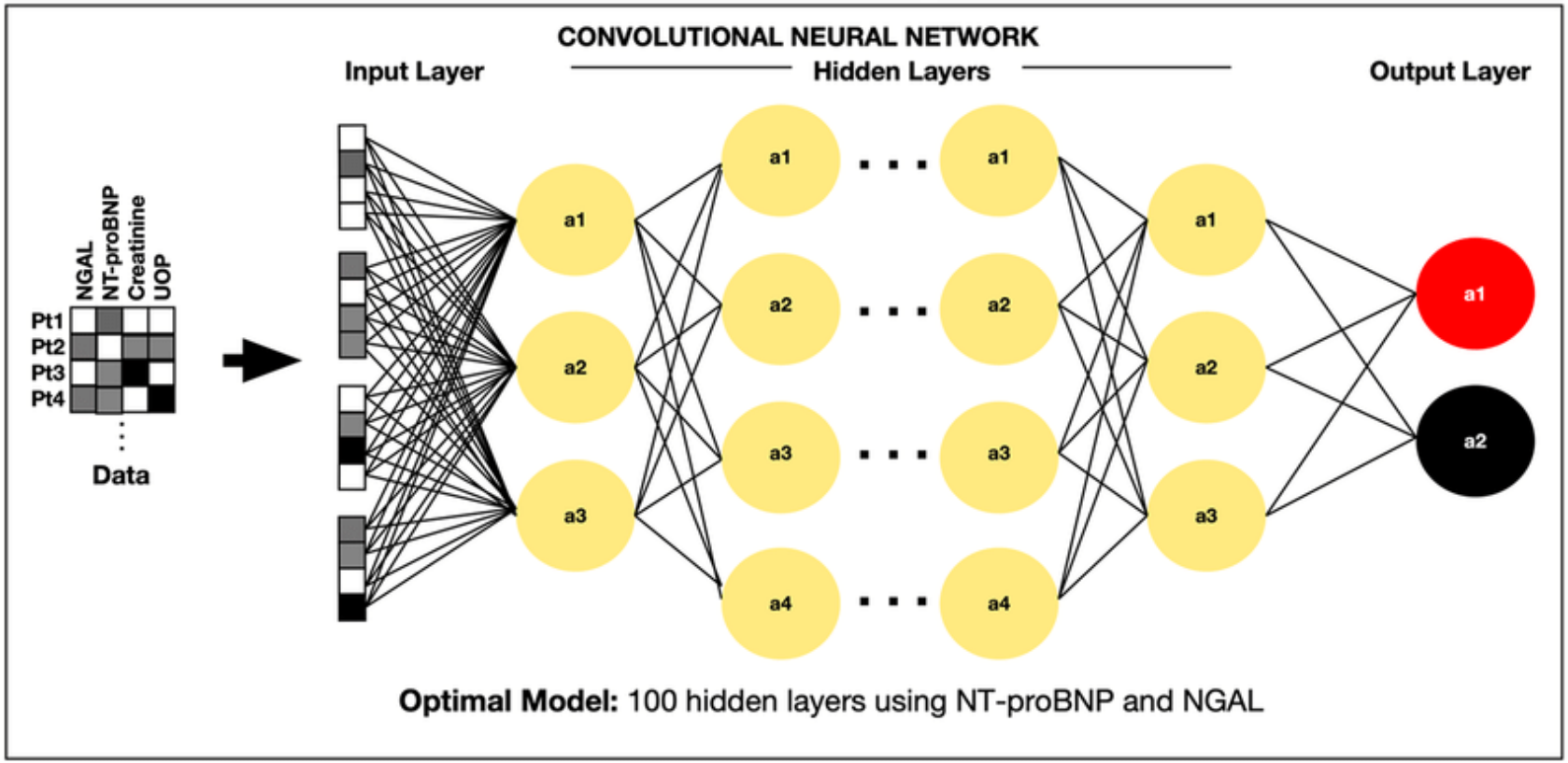
Supervised learning

A plethora of different classification method



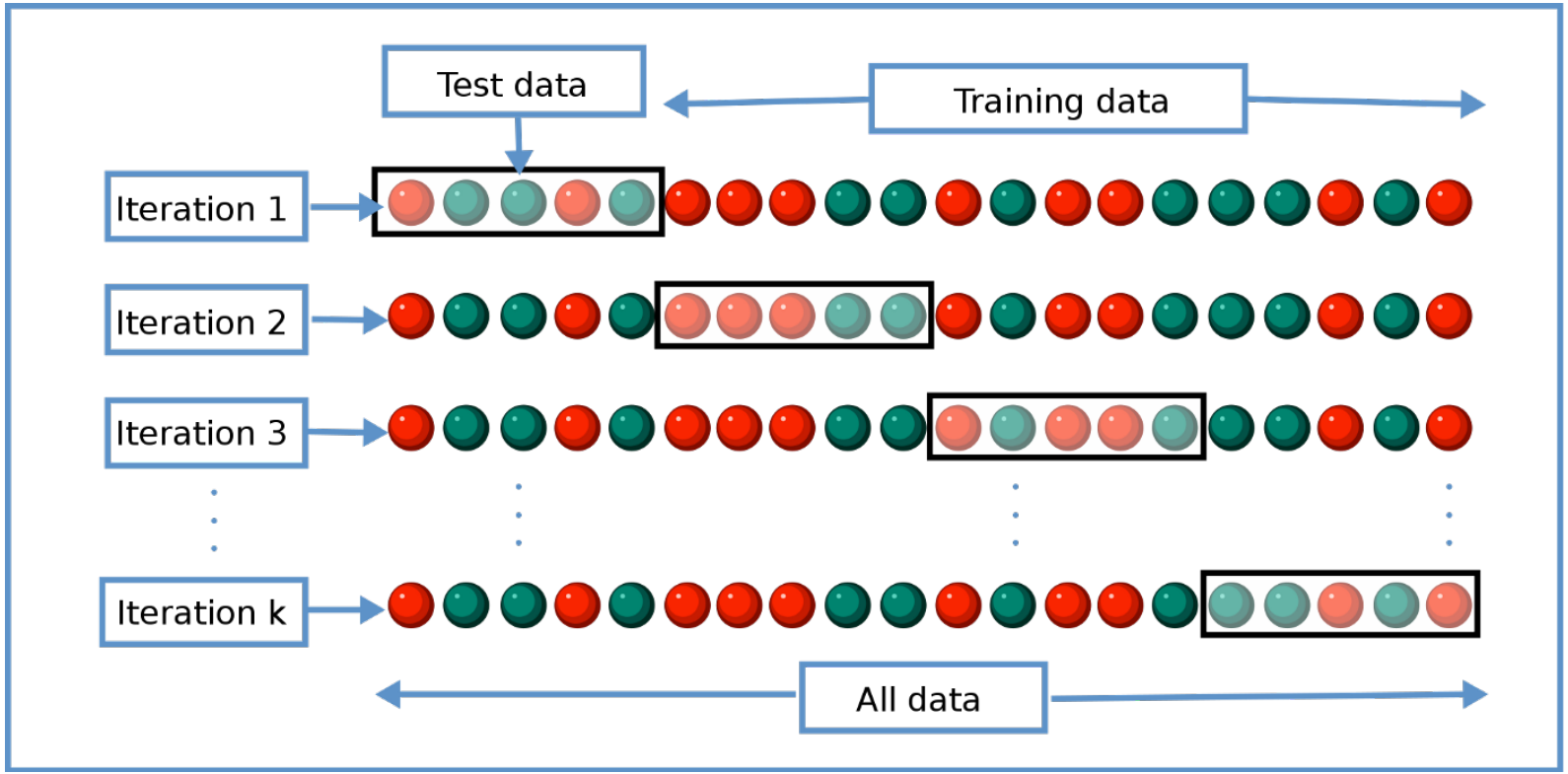
Supervised learning

A plethora of different classification method



Supervised learning

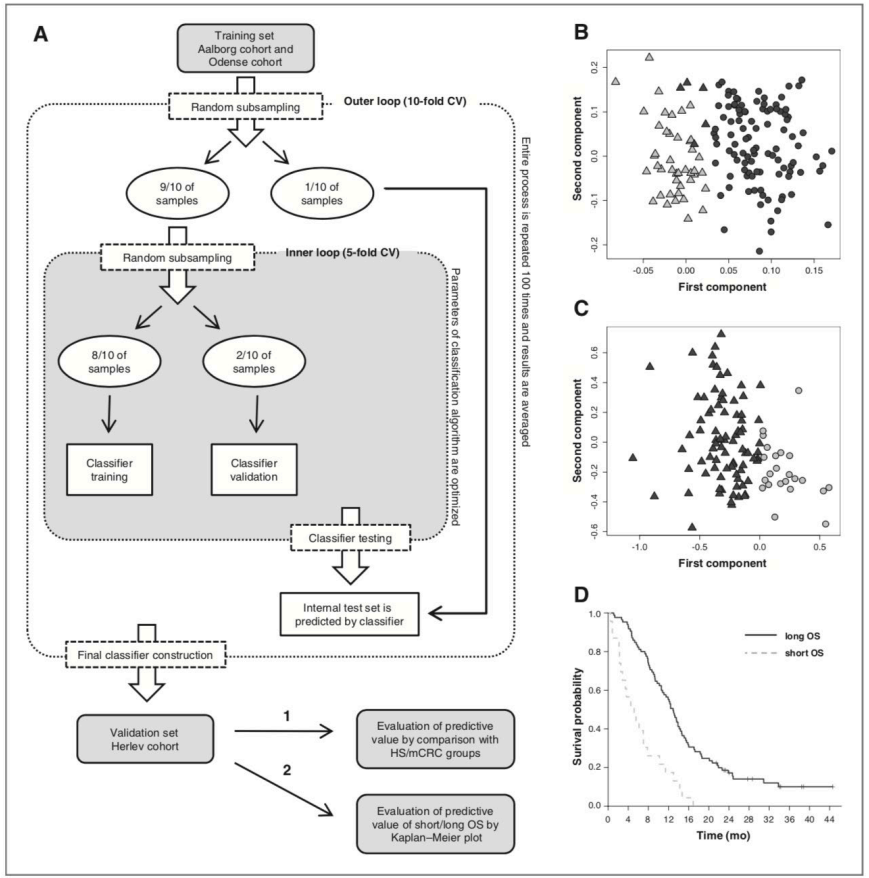
Cross-validation



Supervised learning

Double Cross-validation

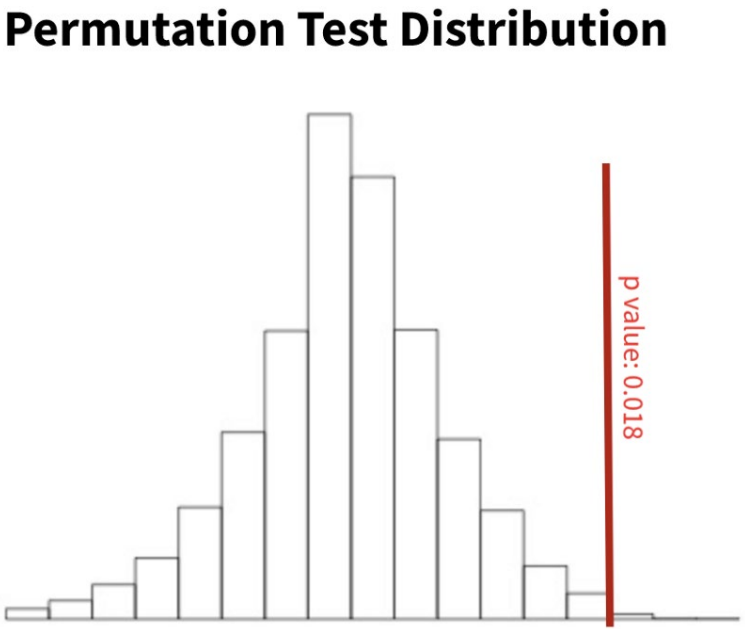
Never adjust your parameters based on the results!



Supervised learning

Permutation test

Permutation tests are non-parametric tests that require very few assumptions. So, when you don't know much about your data generating mechanism (the population), permutation tests are an effective way to determine statistical significance.





ICGEB TRIESTE
ITALY



ICGEB NEW DELHI
INDIA



ICGEB CAPE TOWN
SOUTH AFRICA

ICGEB

International Centre for Genetic Engineering and Biotechnology

Developing Knowledge

Univariate analysis

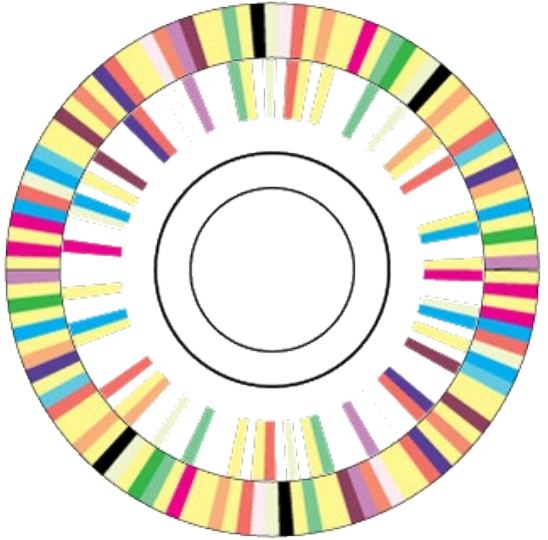
25 August 2021

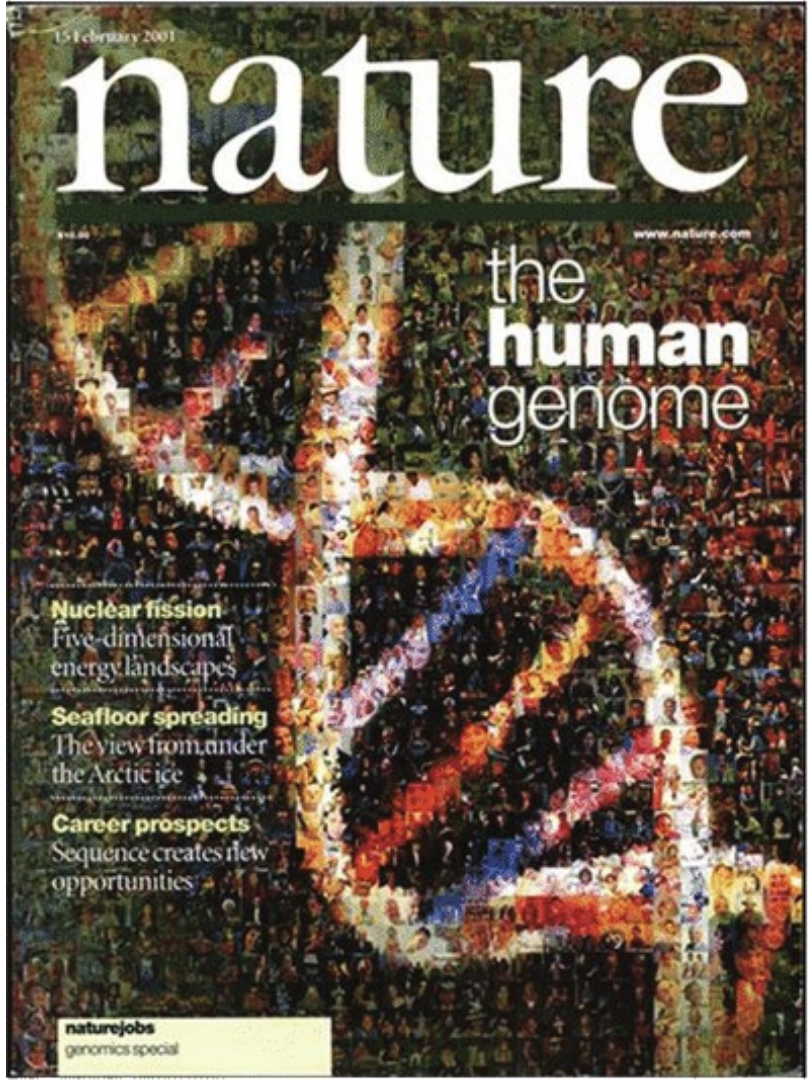
www.icgeb.org

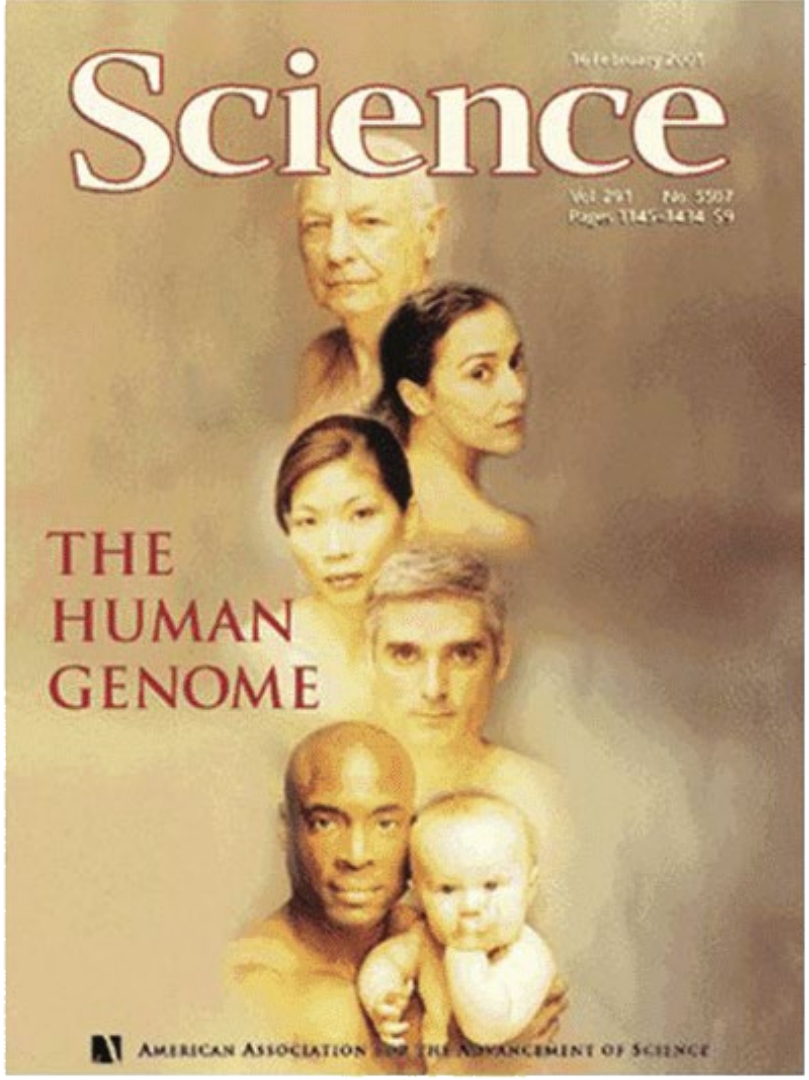
First revolution

Haemophilus influenzae became the first bacterium bacterium genome to be sequenced

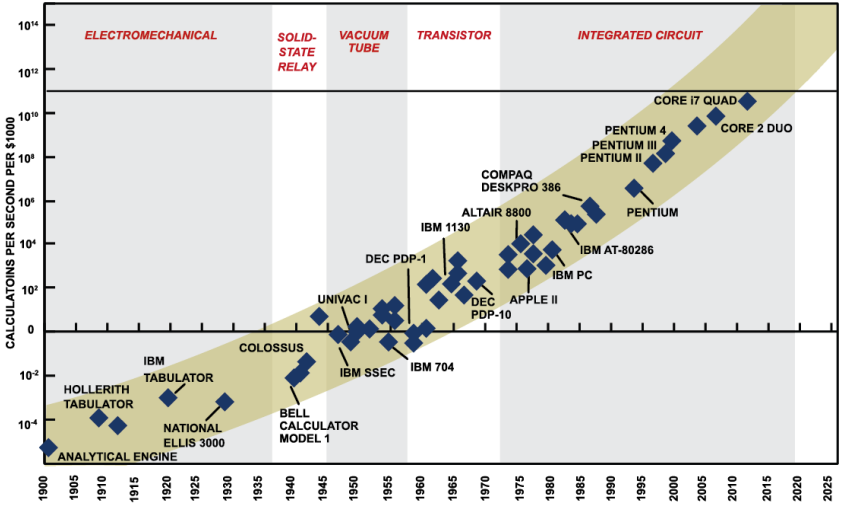
Sequencing the genome of bacterium *Haemophilus influenzae*, reported in **May 1995**, demonstrated for the first time that random "shotgun" sequencing could be applied to whole genomes with speed and accuracy. In conventional sequencing, genomes were laboriously broken down into ordered overlapping segments before sequencing efforts began. In contrast, shotgun sequencing used powerful computational software to assemble shot overlapping DNA fragments, without any preliminary physical map of the genome. Within months after completion of the *H. influenzae* project, the same method was successfully applied to another bacterium, *Mycoplasma genitalium*.



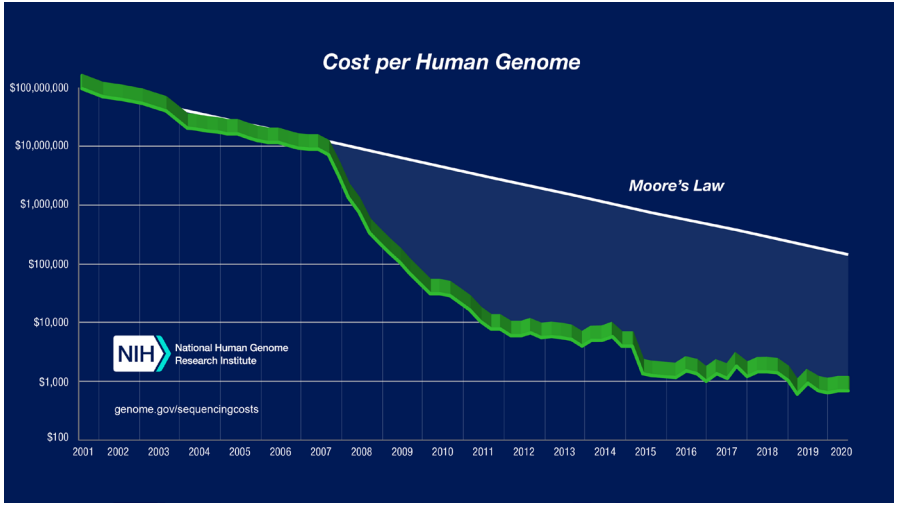

15 February 2001
nature
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special


16 February 2001
Science
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

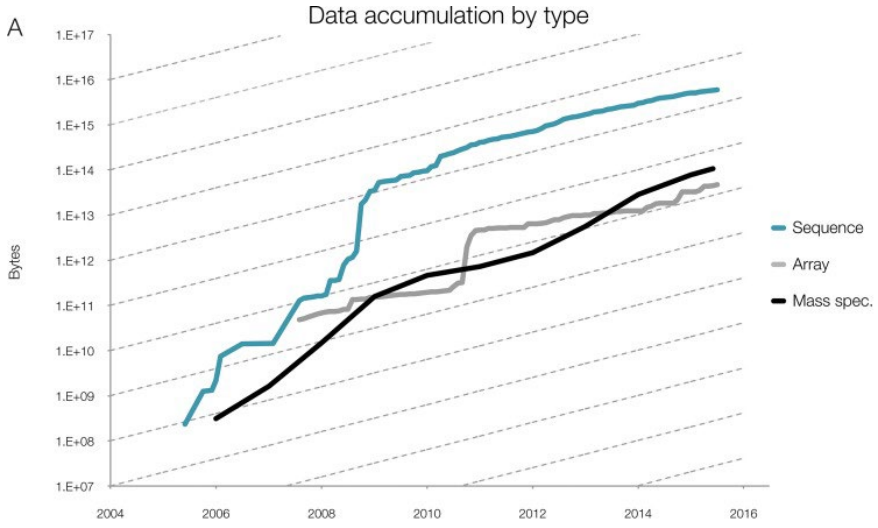
Second revolution

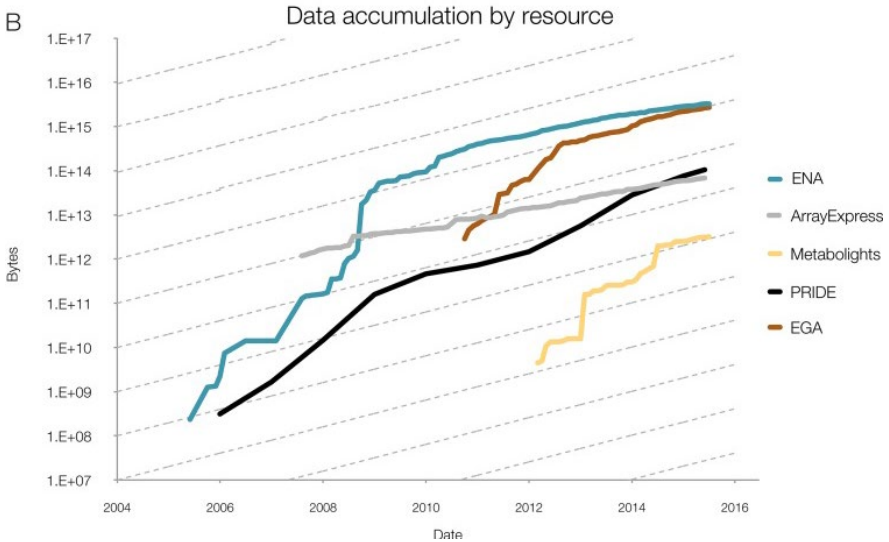


SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.



Third revolution





Data accumulation at EMBL-EBI

Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Res. 2016 Jan 4;44(D1):D20-6.

TCGA datasets

TCGA BY THE NUMBERS

TCGA data describes **33** DIFFERENT TUMOR TYPES ...including **10** RARE CANCERS

...based on paired tumor and normal tissue sets collected from **11,000** PATIENTS

...using **7** DIFFERENT DATA TYPES



RESULTS & FINDINGS

MOLECULAR BASIS OF CANCER	Improved our understanding of the genomic underpinnings of cancer	For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.
TUMOR SUBTYPES	Revolutionized how cancer is classified	TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*
THERAPEUTIC TARGETS	Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development	TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

THE TEAM



20 COLLABORATING INSTITUTIONS across the United States and Canada

WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.



*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

Not Secure | firebrowse.org

FIREBROWSE

Search analysis results

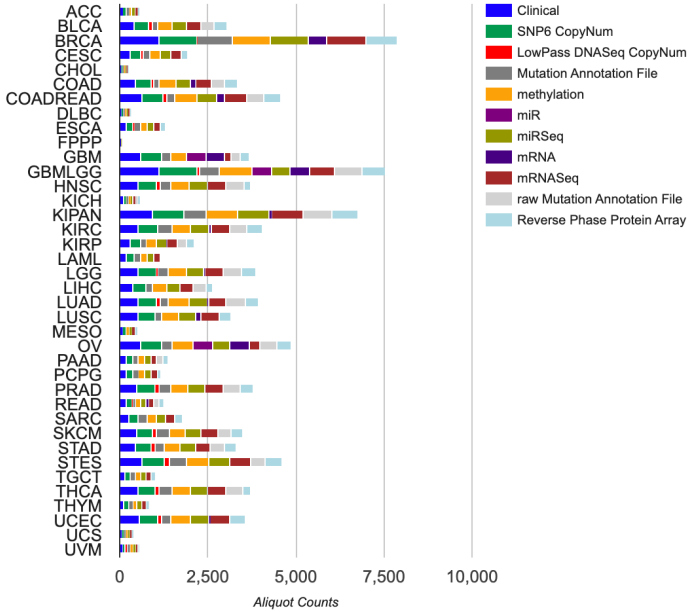
HOME BROAD GDAC WEB API FAQ SAMPLES REPORT AWG RESULTS OLD RUNS TUTORIAL RELEASE NOTES CONTACT

View Expression Profile | Enter gene name | Enter cohort abbrev | View Analysis Profile

SELECT COHORT ▼

- Clinical Analyses
- CopyNumber Analyses
- Correlations Analyses
- miR Analyses
- miRseq Analyses
- mRNA Analyses
- mRNAseq Analyses
- Mutation Analyses
- Pathway Analyses
- RPPA Analyses

TCGA data version 2016_01_28



Not Secure | firebrowse.org

FIREBROWSE

Search analysis results

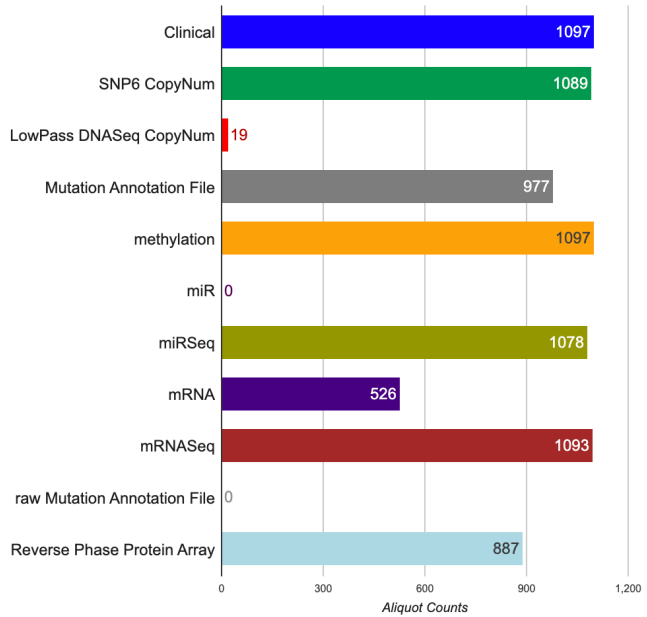
HOME | BROAD GDAC | WEB API | FAQ | SAMPLES REPORT | AWG RESULTS | OLD RUNS | TUTORIAL | RELEASE NOTES | CONTACT

View Expression Profile | Enter gene name | BRCA | View Analysis Profile

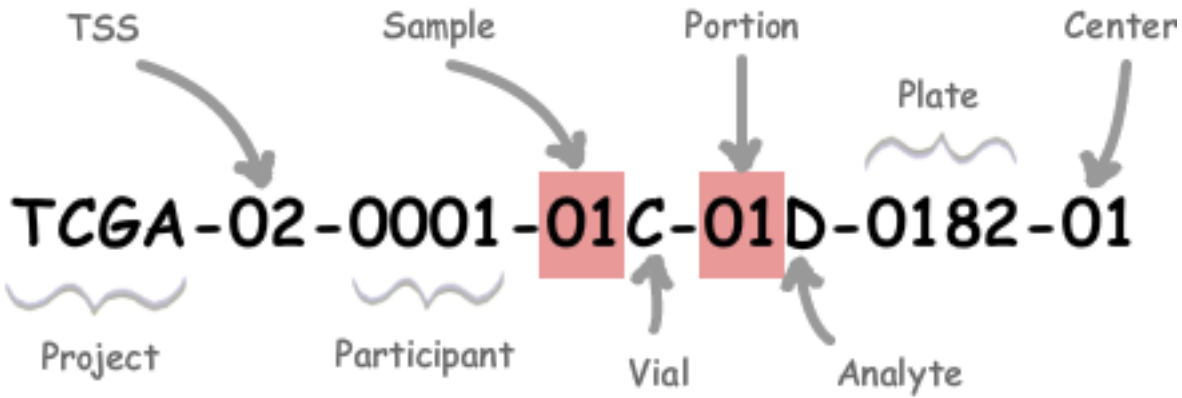
Breast invasive carcinoma (BRCA)

- Clinical Analyses
- CopyNumber Analyses
- Correlations Analyses
- Methylation Analyses
- miRseq Analyses
- mRNA Analyses
- mRNAseq Analyses
- Mutation Analyses
- Pathway Analyses
- RPPA Analyses

TCGA data version 2016_01_28 for BRCA



TCGA code


TSS
Sample
Portion
Plate
Center
TCGA-02-0001-01C-01D-0182-01
Project
Participant
Vial
Analyte