V Jornadas de Bioinformática. Universidad de Granada

Integrative Omics Data Analysis, Lights and Shadows

Alex Sánchez–Pla

February 25, 2021

Genetics, Microbiology & Statistics
Facultad de Biología, Universitat de Barcelona
Statistics and Bioinformatics Unit (UEB)
Vall Hebron Institut de Recerca

Table of Contents

Presentation and Objectives

Who, where, what?

Statistics and Bioinformatics
Integrative analysis of omics data

Nutrition and Metabolomics

Software development

Alex Sánchez

Associate Professor of Statistics.
Faculty of Biology Universitat de Barcelona
UB Director MSc of Statistics & Bioinformatics

Head of *Data Science Projects* Platform

Vall d'Hebron Institut de Recerca

Multiversity teaching

Análisis de Datos Omicos (>300 alum)

Omics Data Analysis (< 10 alum)

Omics Techniques (~40 alum)

The UOC-UB MSc in Bostatistics and Bioinformatics

Universitat Oberta de Catalunya

Estudia en la UOC

Estudios por titulación | Estudios por área | Acceso y matrícula | ¿Por qué la UOC? | Campus

Másters universitarios

Máster universitario de Bioinformática y Bioestadística (interuniversitario: UOC, UB)

Presentación

Presentación
Plan de estudios
Itinerario académico
Objetivos, perfiles y competencias
Requisitos de acceso
Reconocimiento de créditos
Salidas profesionales
Equipo docente
Matrícula
Precios, becas y descuentos

El máster universitario de Bioinformática y Bioestadística en línea de la UOC y la UB forma a **profesionales expertos** en el uso de la tecnología para la gestión, el **análisis y la interpretación de datos biológicos y médicos**.

Créditos: 60 ECTS
Título: Bioinformática y Bioestadística (interuniversitario: UOC, UB)
Inicio: 15 septiembre 2021
Idioma: Castellano, Catalán

El **impacto de la tecnología** en disciplinas como la biología molecular, la medicina, la veterinaria o la agronomía conlleva una gran cantidad de **nuevos datos biológicos**, que son información de valor crucial para el futuro de la **investigación, la innovación y el desarrollo**. Disponer de especialistas en el manejo, el análisis y la interpretación de estos datos es imprescindible. Para ello se necesitan **profesionales de bioinformática**, **genómica**, **biología computacional** o **biocomputación**.

En este contexto, la Universitat Oberta de Catalunya (**UOC**) y la Universidad de Barcelona (**UB**) han unido su **experiencia, su prestigio y sus conocimientos** en el máster universitario de Bioinformática y Bioestadística, el único en España que incorpora una **visión combinada** de ambas disciplinas.

Se trata de una **titulación oficial** con una **formación de calidad y rigor académico** que permite al estudiante adquirir un **perfil profesional**

Próxima matrícula: abril 2021

Información de precio y matrícula

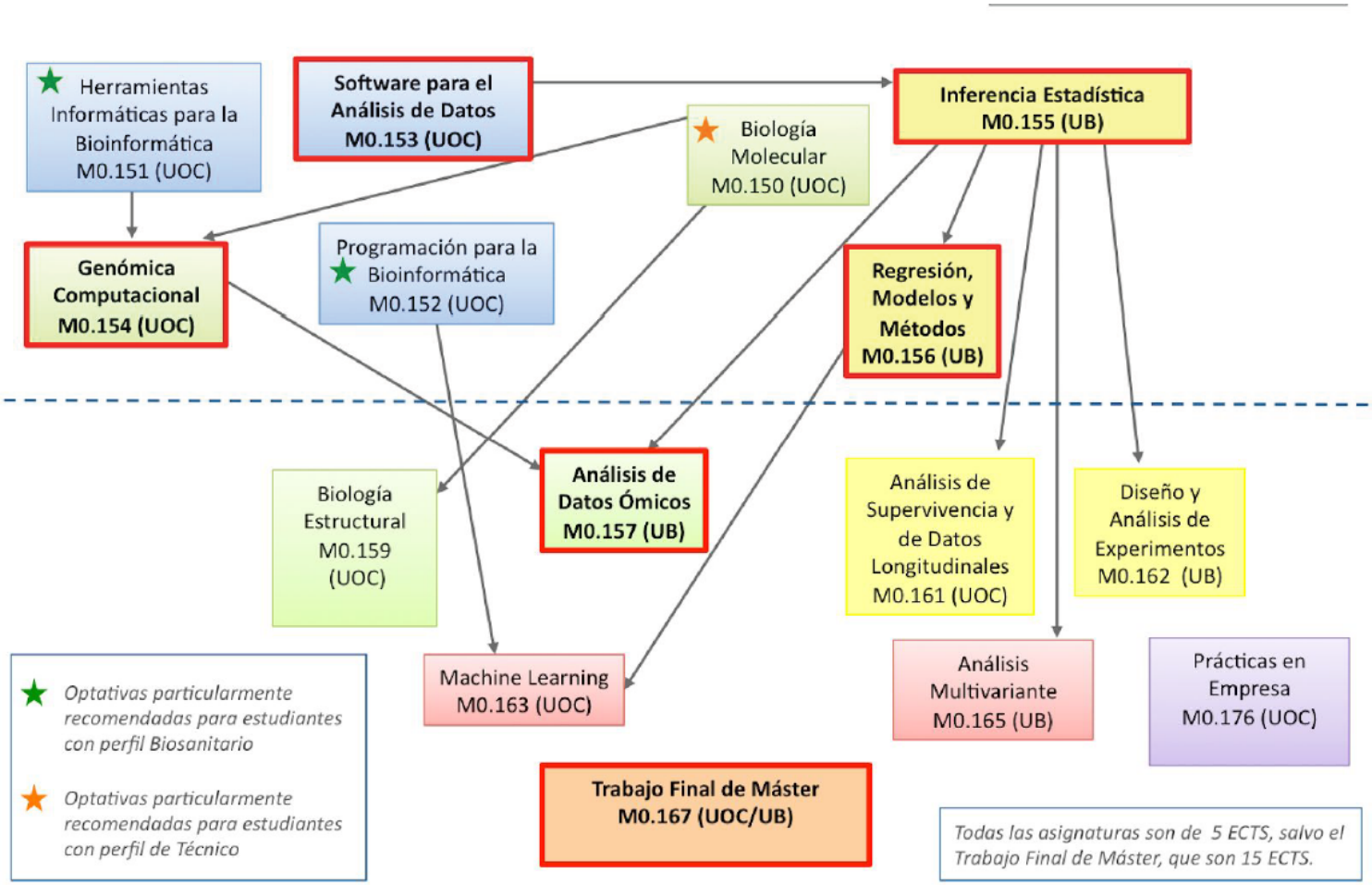
¿Quieres más información?

Envía tus datos y recibirás información de este programa y relativa a productos, servicios y actividades promocionales de la UOC

Nombre
Primer apellido
Segundo apellido

Formamos

The UOC-UB MSc in Bostatistics and Bioinformatics

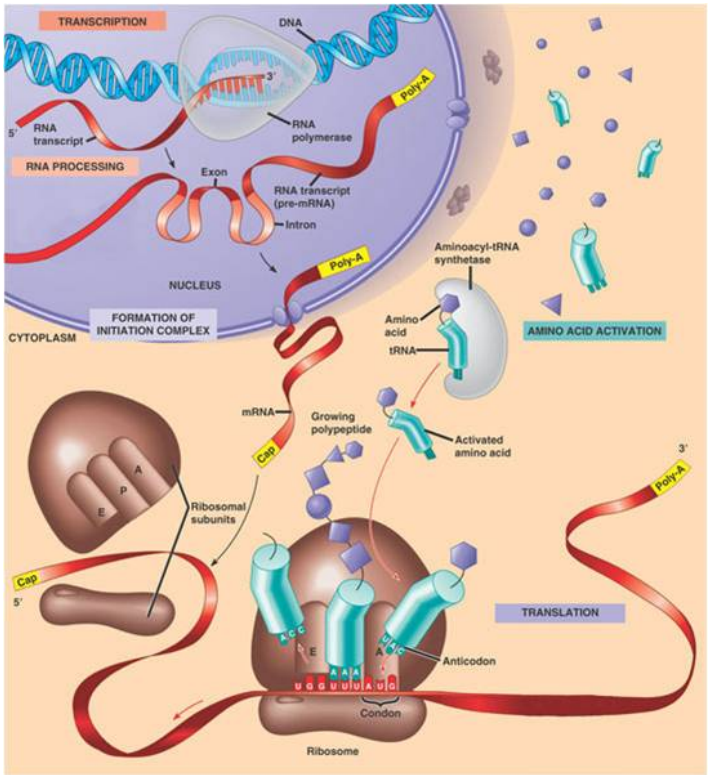


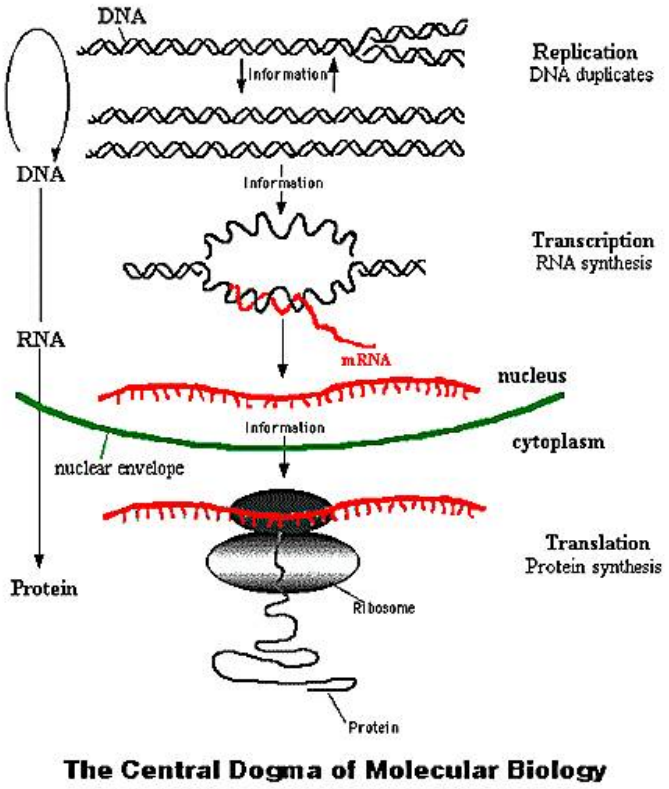
Omics, Omics Data and Omics Integration

What is "omics"?

- In biological context , the suffix "omics" is used to refer to the study of large sets of biological molecules (Smith et al., 2005)
- The study of different components participating and/or regulating complex biological processes, triggered the development of several fields that, together, are described with the term OMICS.

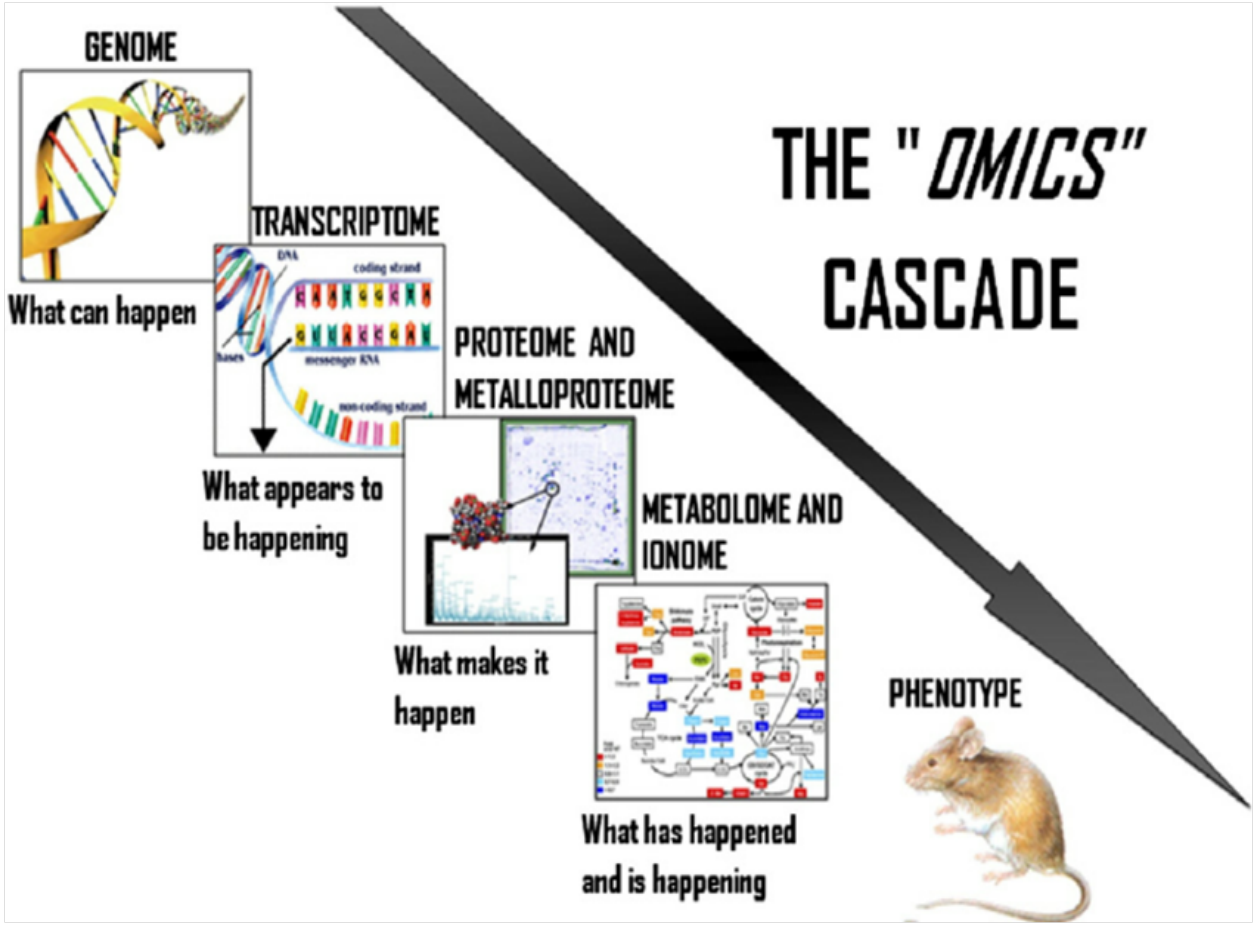
The central dogma



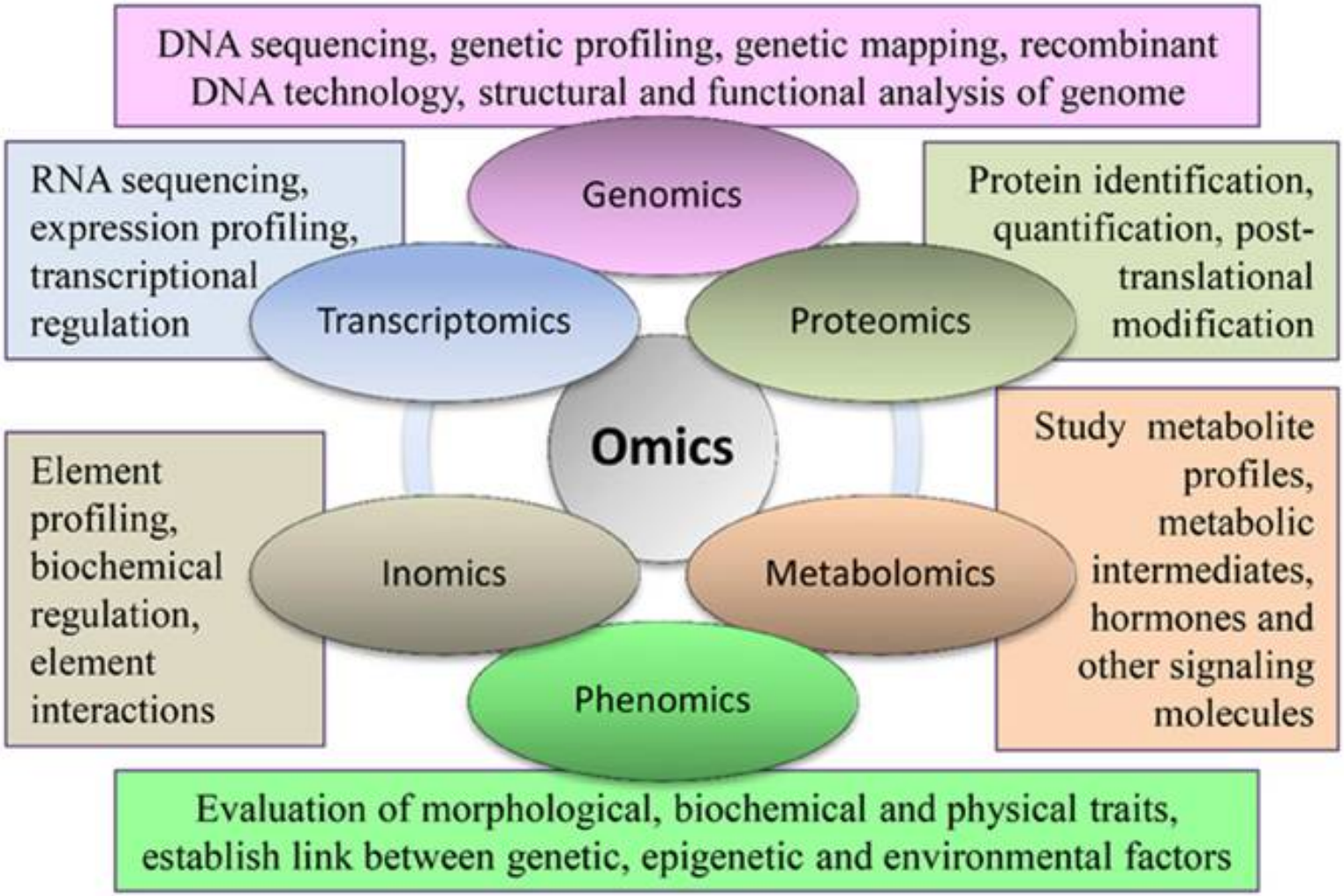


The Central Dogma of Molecular Biology

The Omics Cascade (1): "omes"

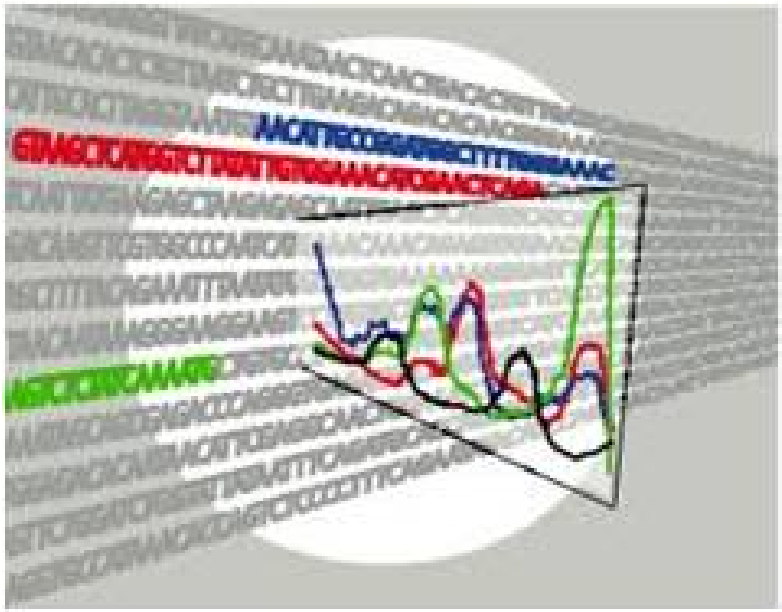


We study omes with omics



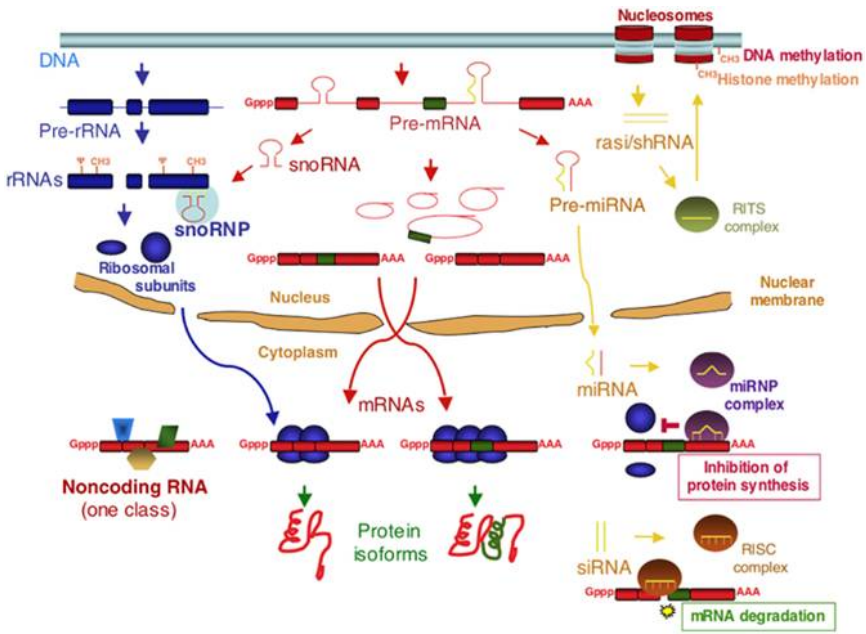
Genomics

Genomics is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes (the complete set of DNA within a single cell of an organism)



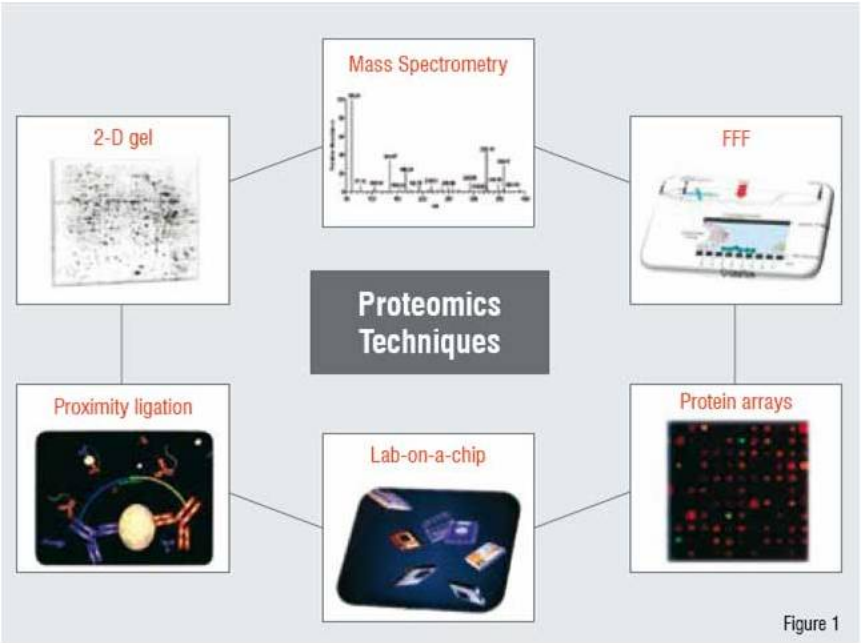
Transcriptomics

- The transcriptome is the set of all RNA molecules, in one or a population of cells.
- Transcriptomics, examines expression levels of mRNAs in a given cell population, often using high-throughput techniques: microarrays or NGS.



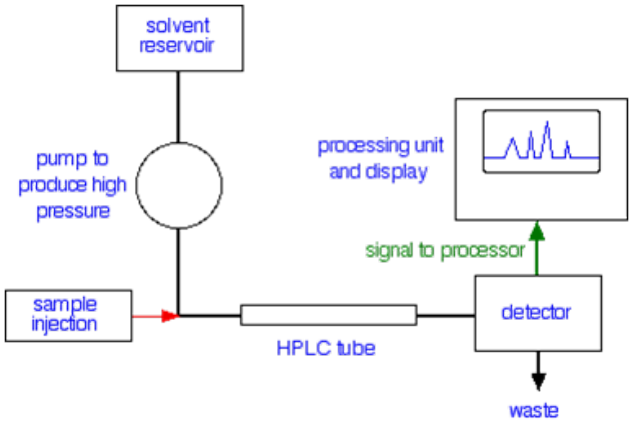
Proteomics

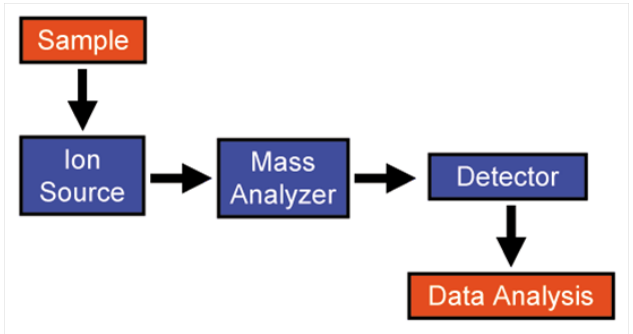
- The large–scale study of proteins (the proteome), particularly their structure and function.
- Relies on a wide spectra of techniques
  - 2D gel based
  - Mass Spectrometry (MS)
  - Seldi-TOF (MS)
  - Protein Arrays
  - ...



Metabolomics

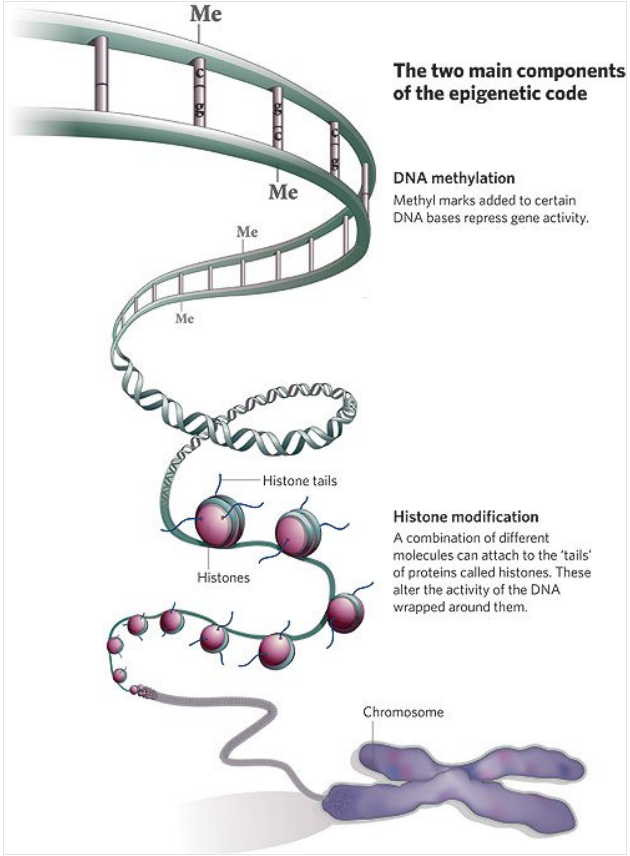
- Comprehensive and simultaneous systematic determination of
  - metabolite levels in the metabolome and
  - their changes over time as a consequence of stimuli.
- Relies on
  - Separation techniques: GC, CE, HPLC, UPLC
  - Detection techniques: NMR, MS



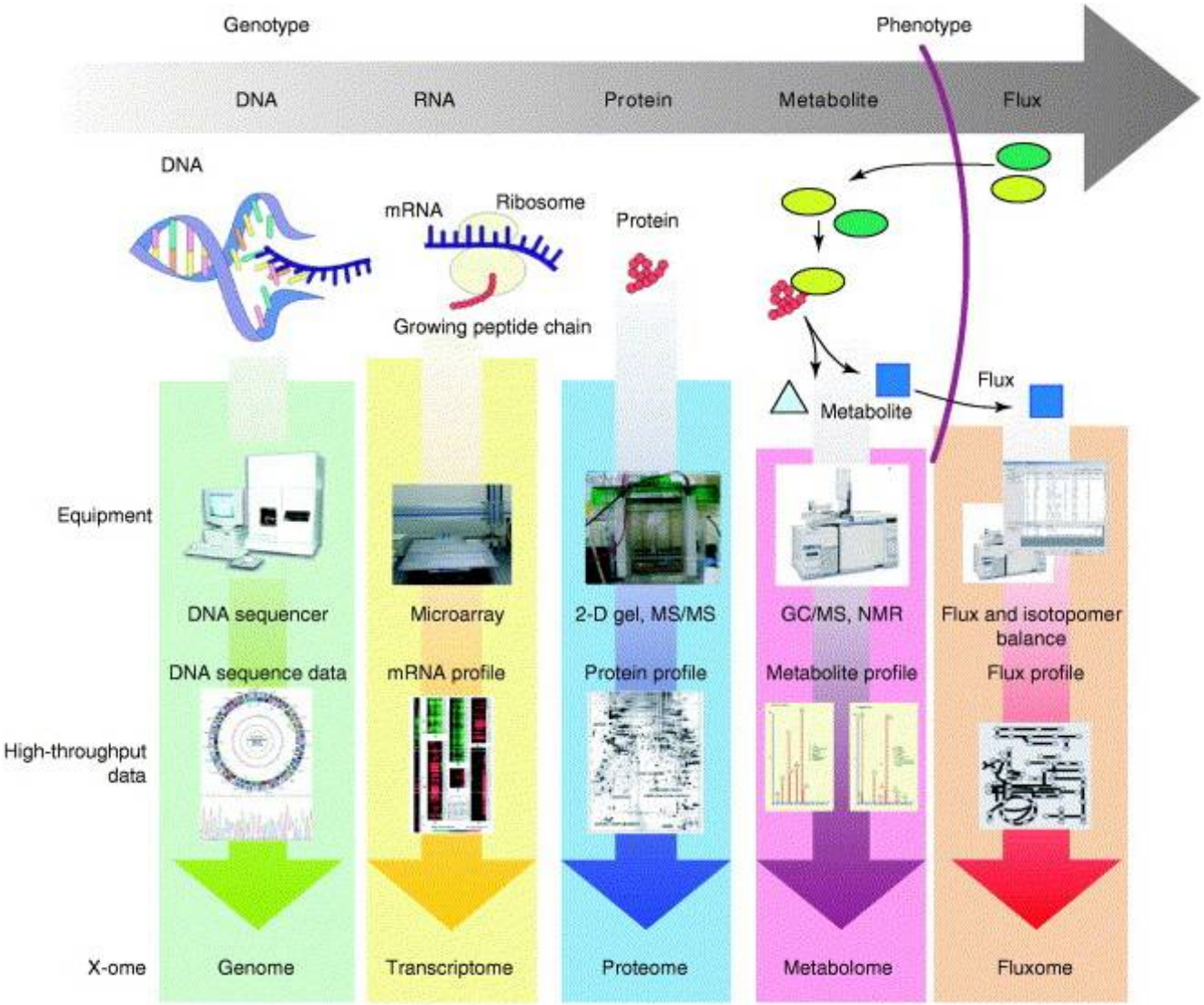


Epigenetics and Epigenomics

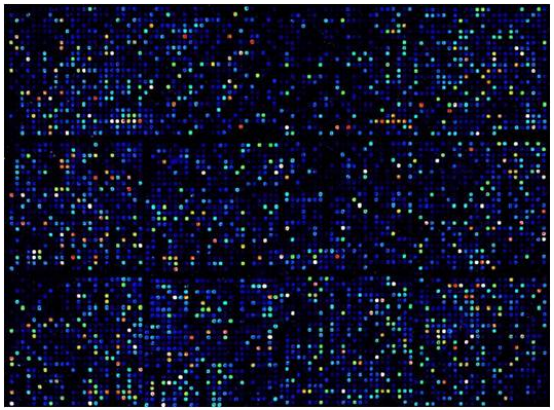
- *Epigenetics* studies changes in the phenotype or gene expression caused by other mechanisms than changes in the underlying DNA sequence.
  - DNA methylation
  - Histone modifications
- Epigenetics refers to the study of single genes or sets of genes. Epigenomics refers to global analyses of epigenetic changes across the entire genome

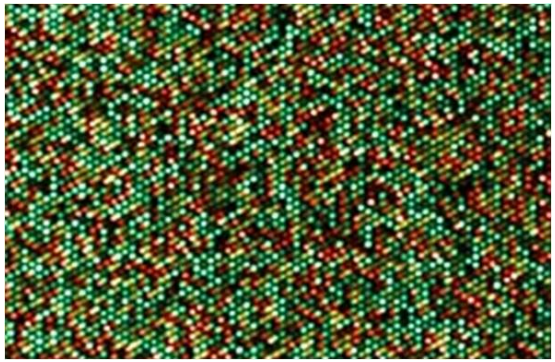


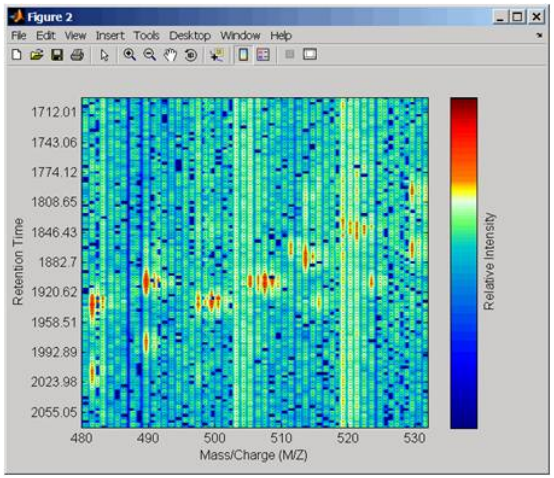
In summary we use "omics" to study "omes"

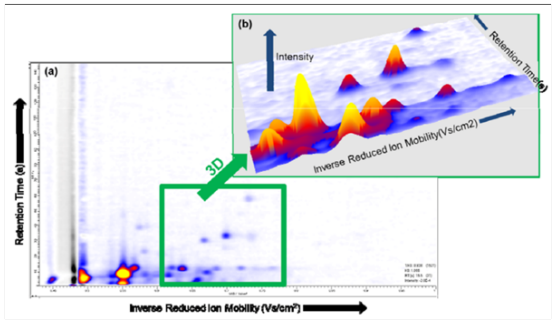


Omics data are high throughput





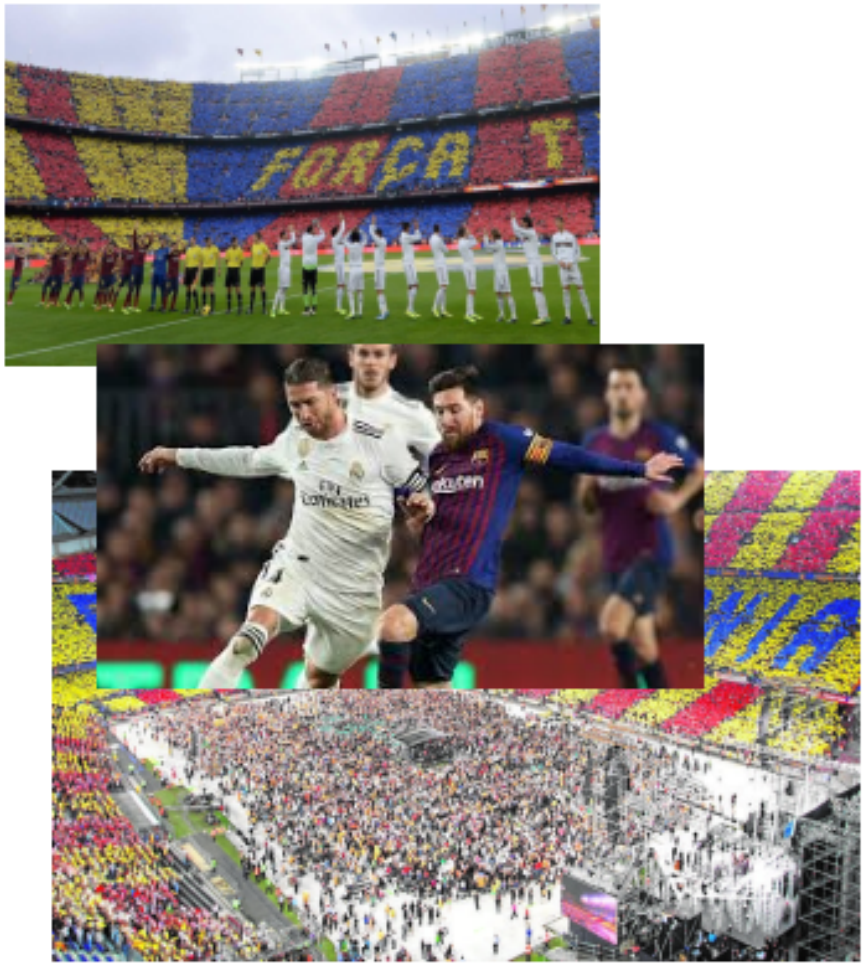


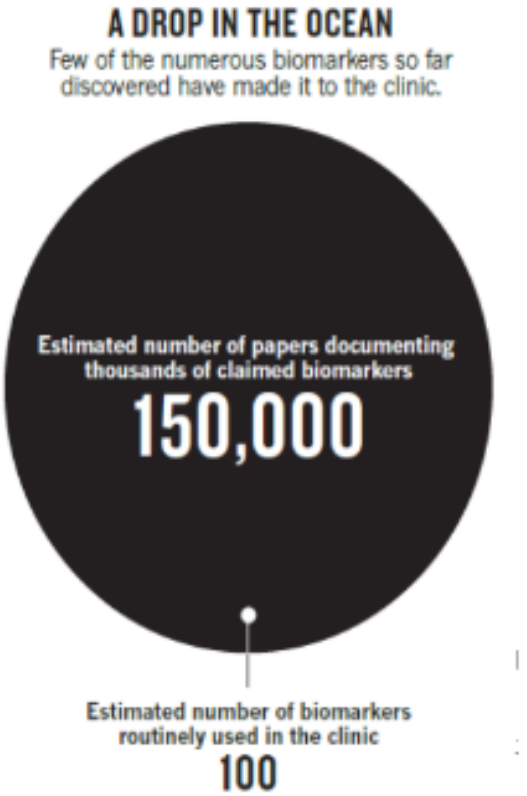


Bioinformatics and Biostatistics are essential

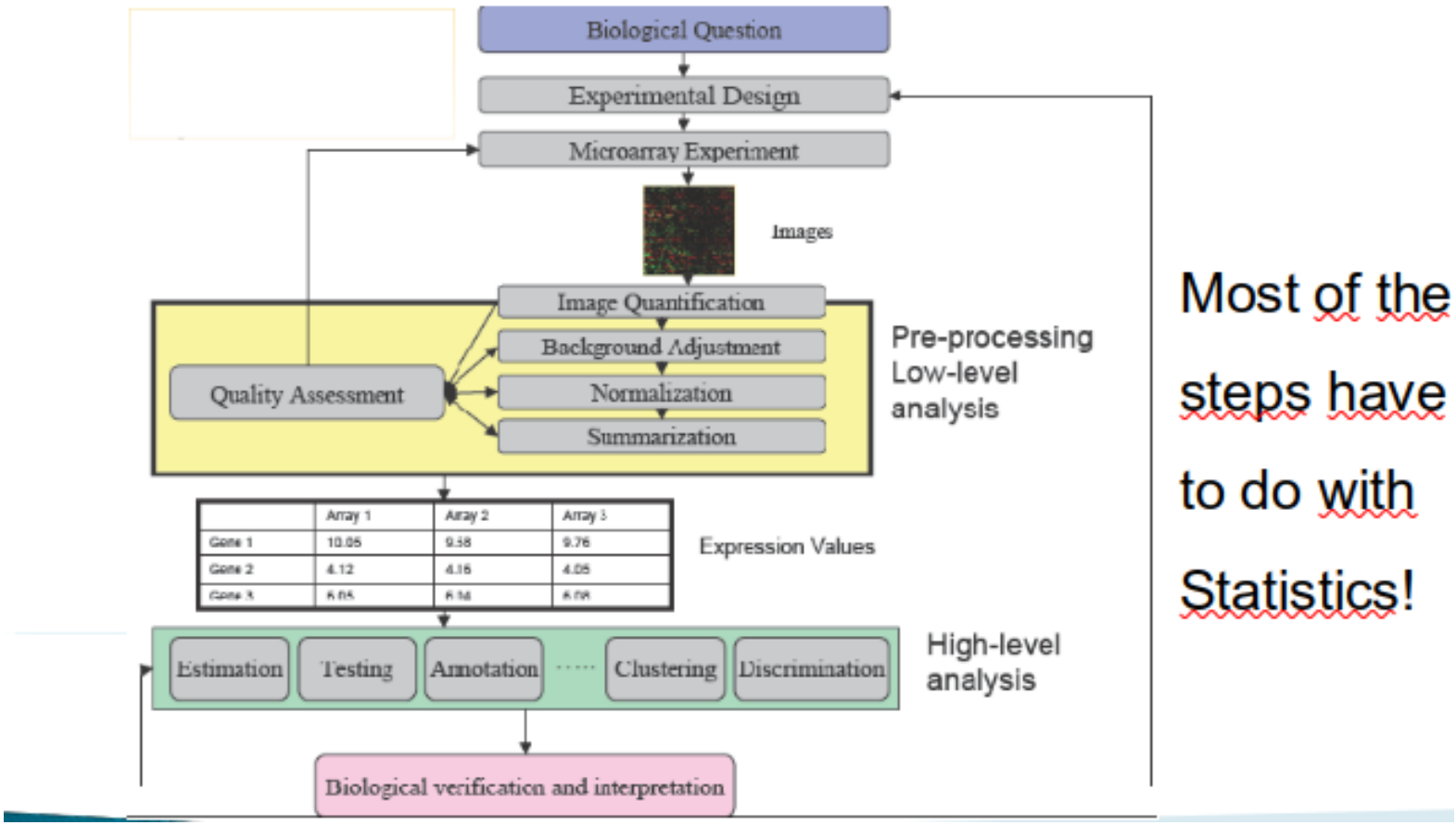
"Data don't make any sense,
we will have to resort to statistics."

Not to talk of noise

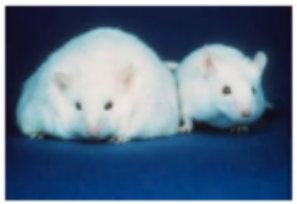


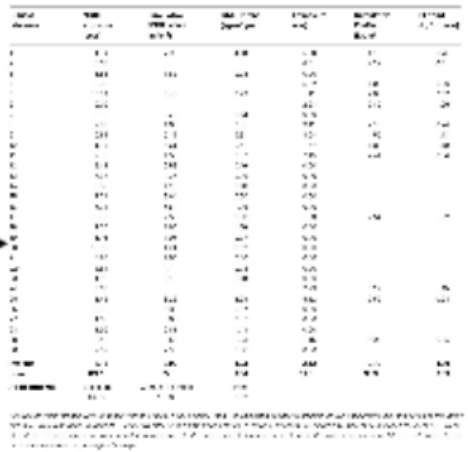


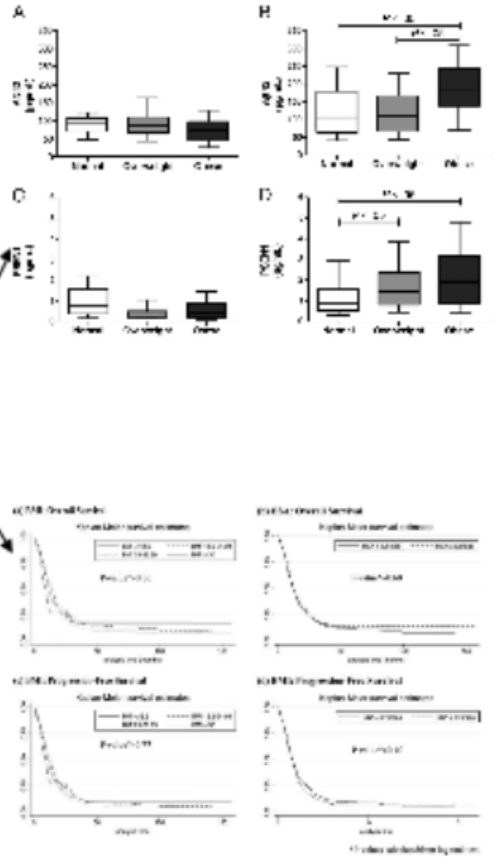
How do we analyze these data



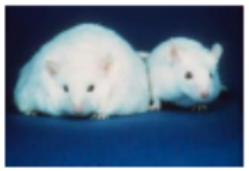
How we studied disease in the 20th century

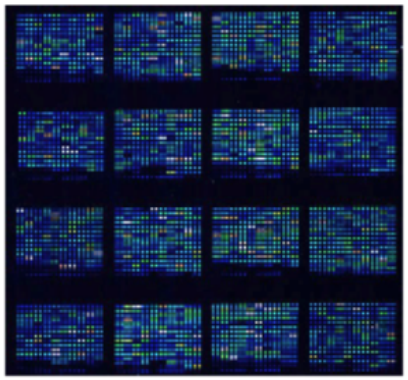


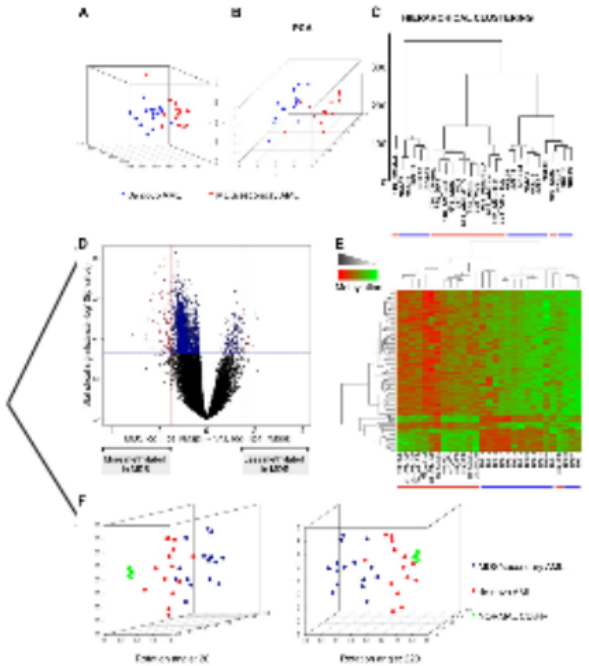




Single-omics analysis: The first decade of the XXIst

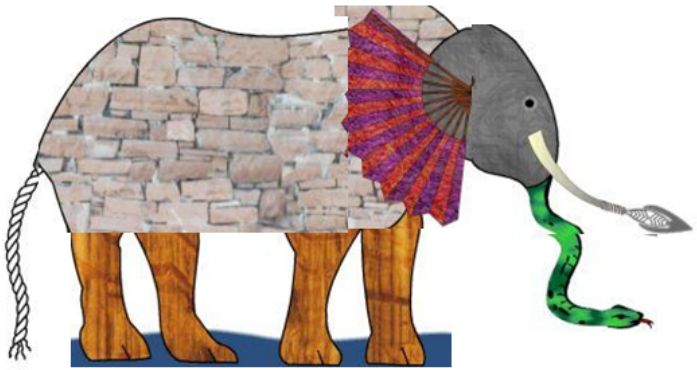






Why should we integrate data?

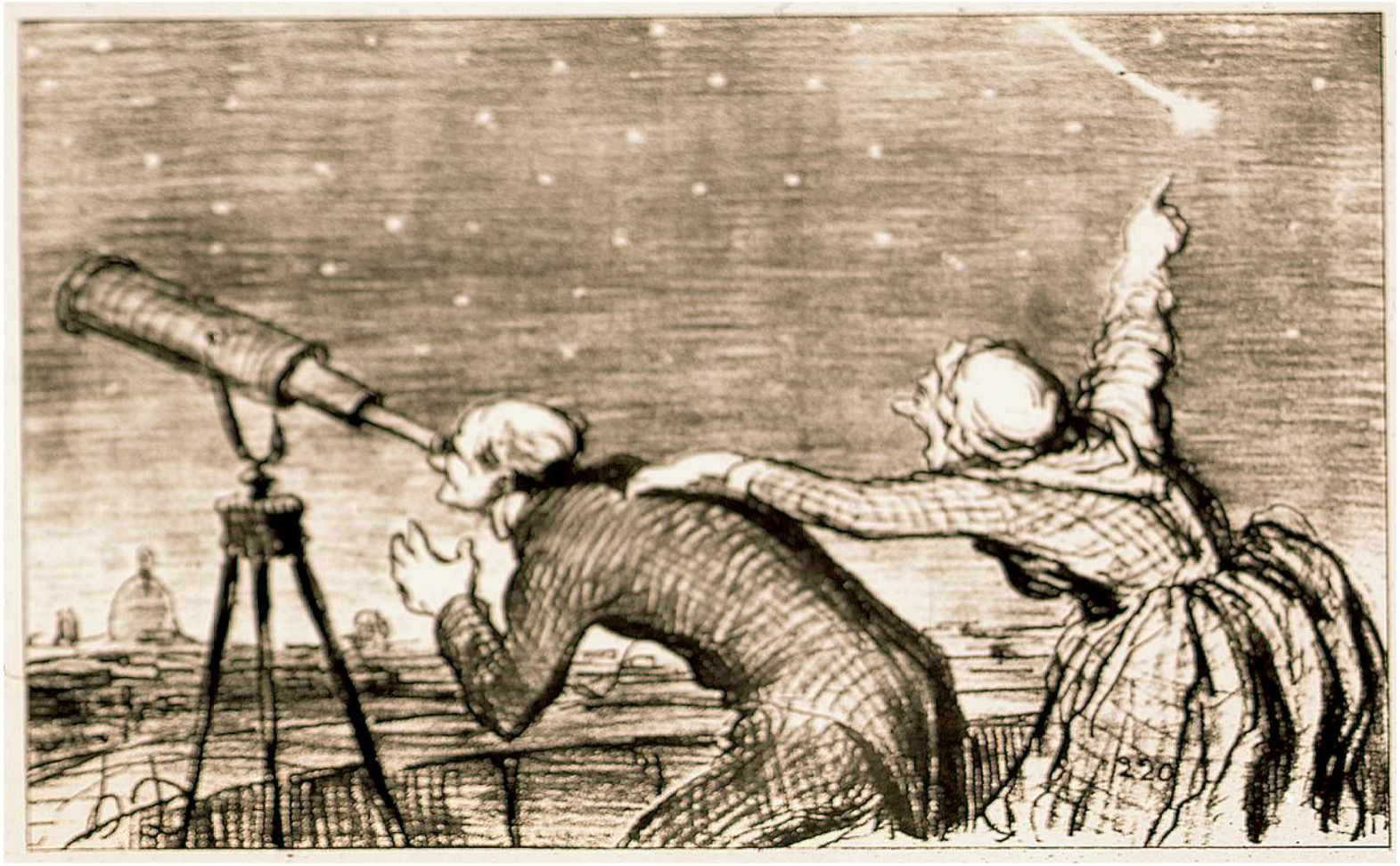
The Blind Men and the Elephant



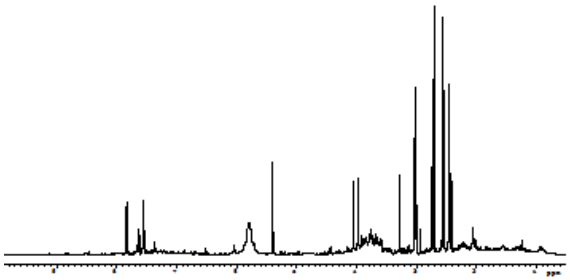
What we learn from an experiment may depend on where we look, how we look, and the scope of our view!

`http://www.noogenesis.com/pineapple/blind_men_elephant.html`

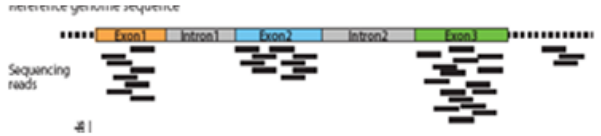
Focussing only on one platform risks missing an obvious signal



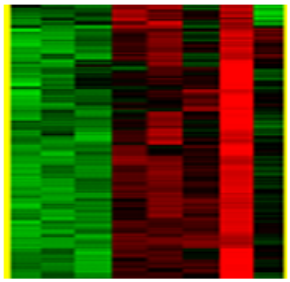
So let's measure as many as possible



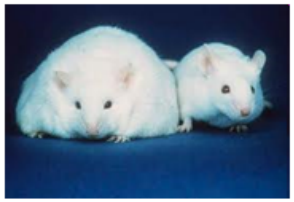
H NMR metabolites

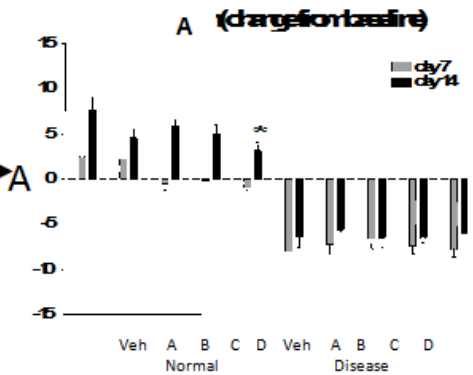


"NGS-Sequences

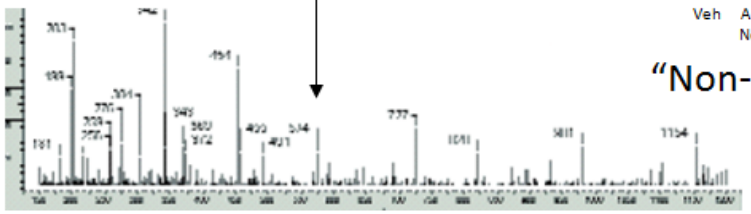


Affy Transcriptome



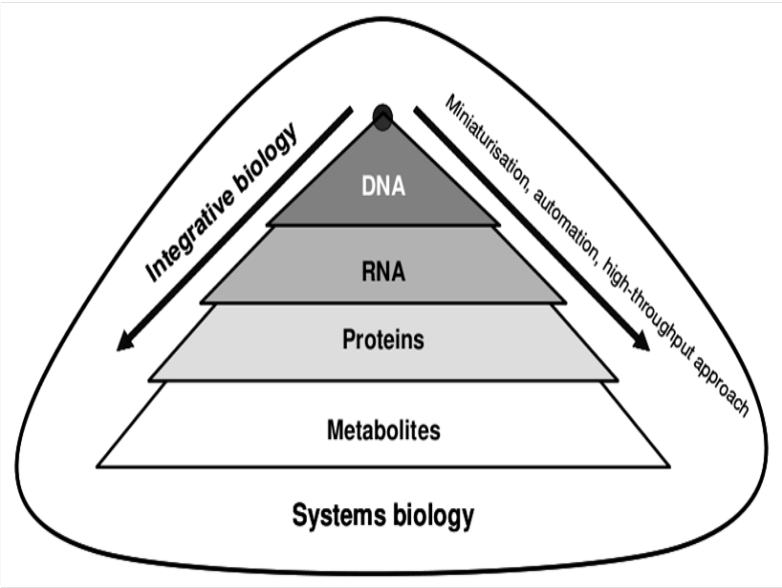


"Non-omic" markers



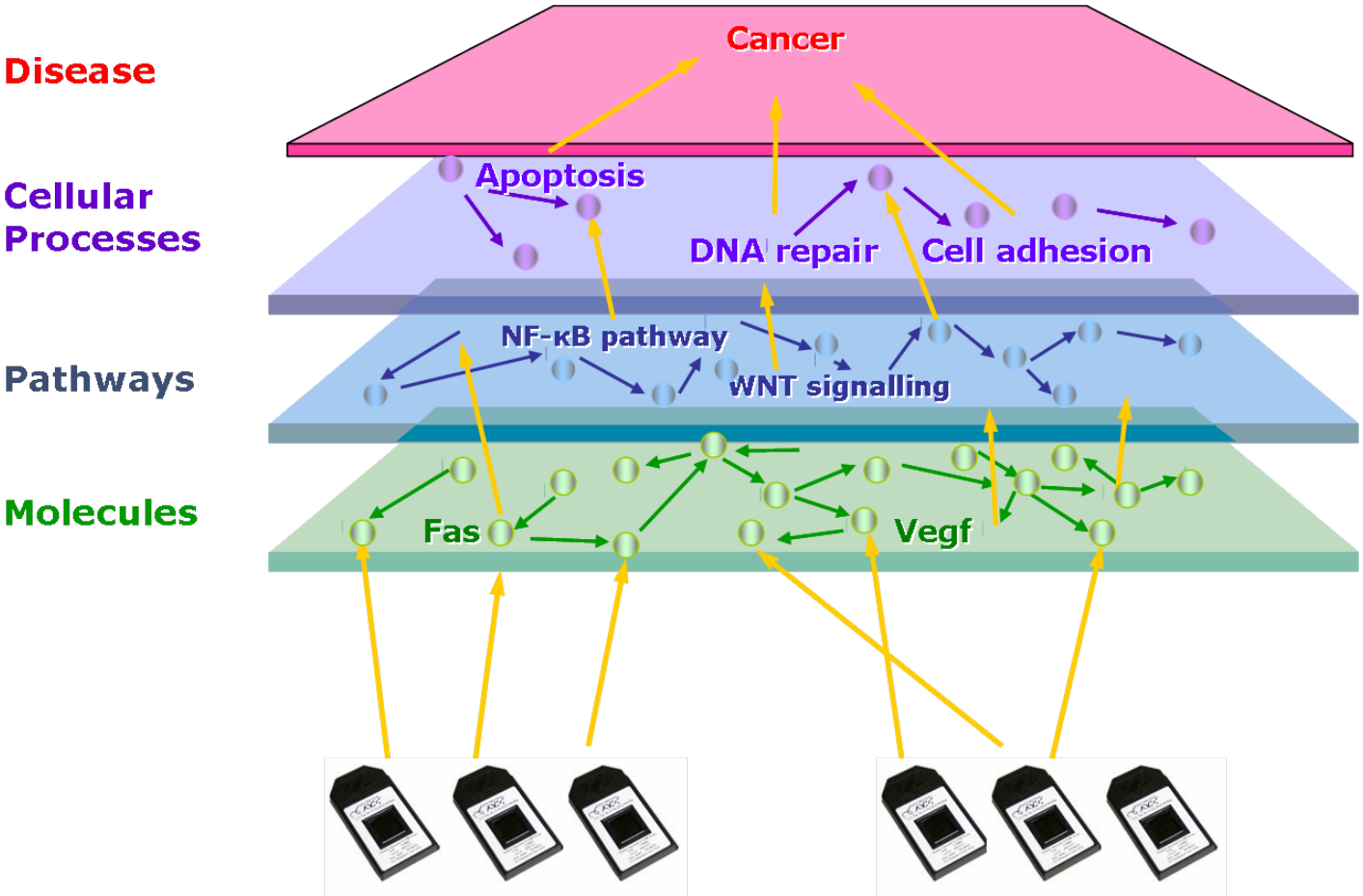
LC-MS proteomicss

From componentwise to global approaches

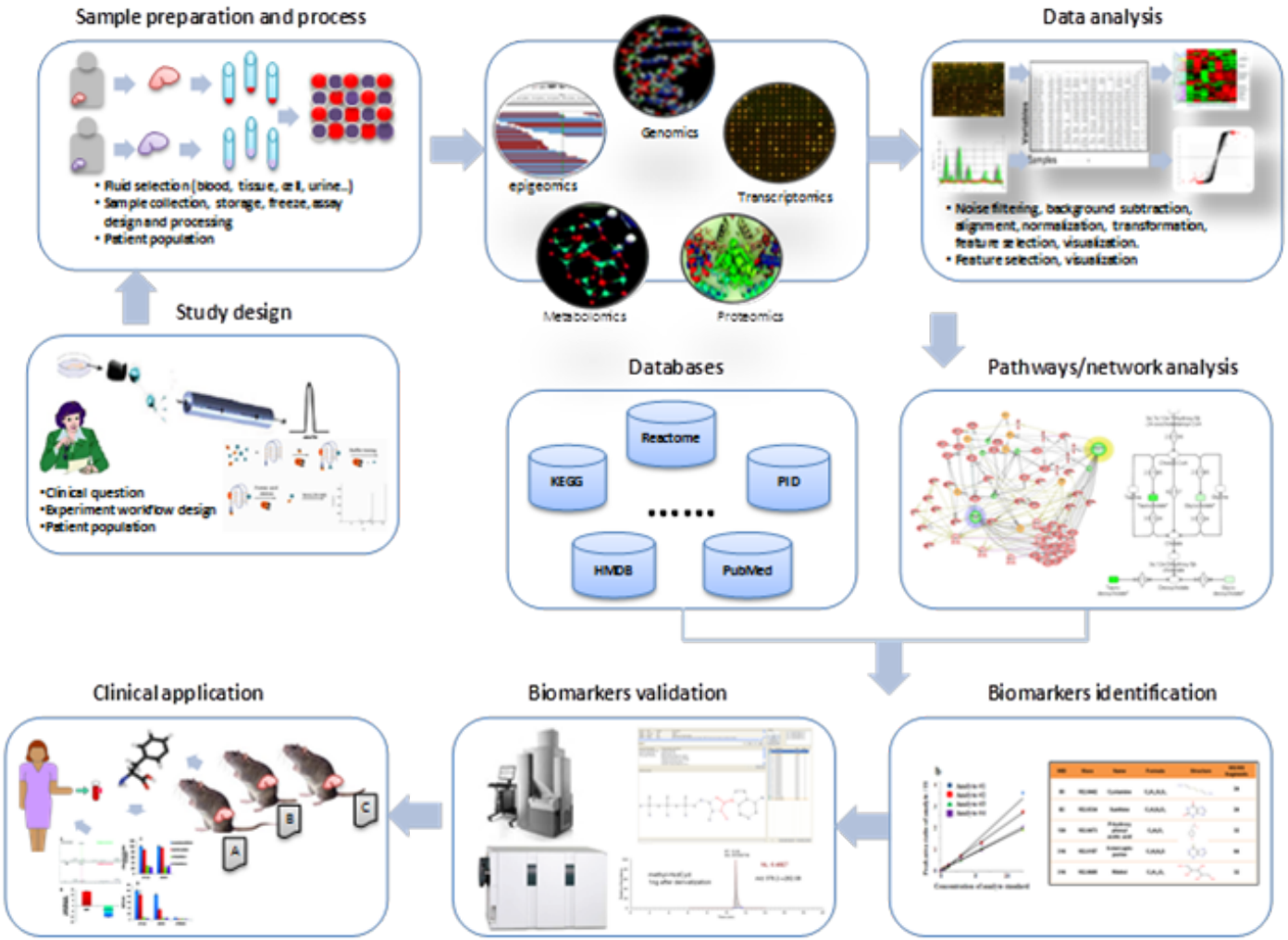


- It is expected that the integrated collection and analysis of diverse types of data,
- jointly modelled and analyzed in a systems biology approach
- can shed light on the global functioning of biological systems.

Ultimate Goal (1): understanding of complex processes



Ultimate Goal (2): Improved Robust Biomarkers



But what is Data Integration?

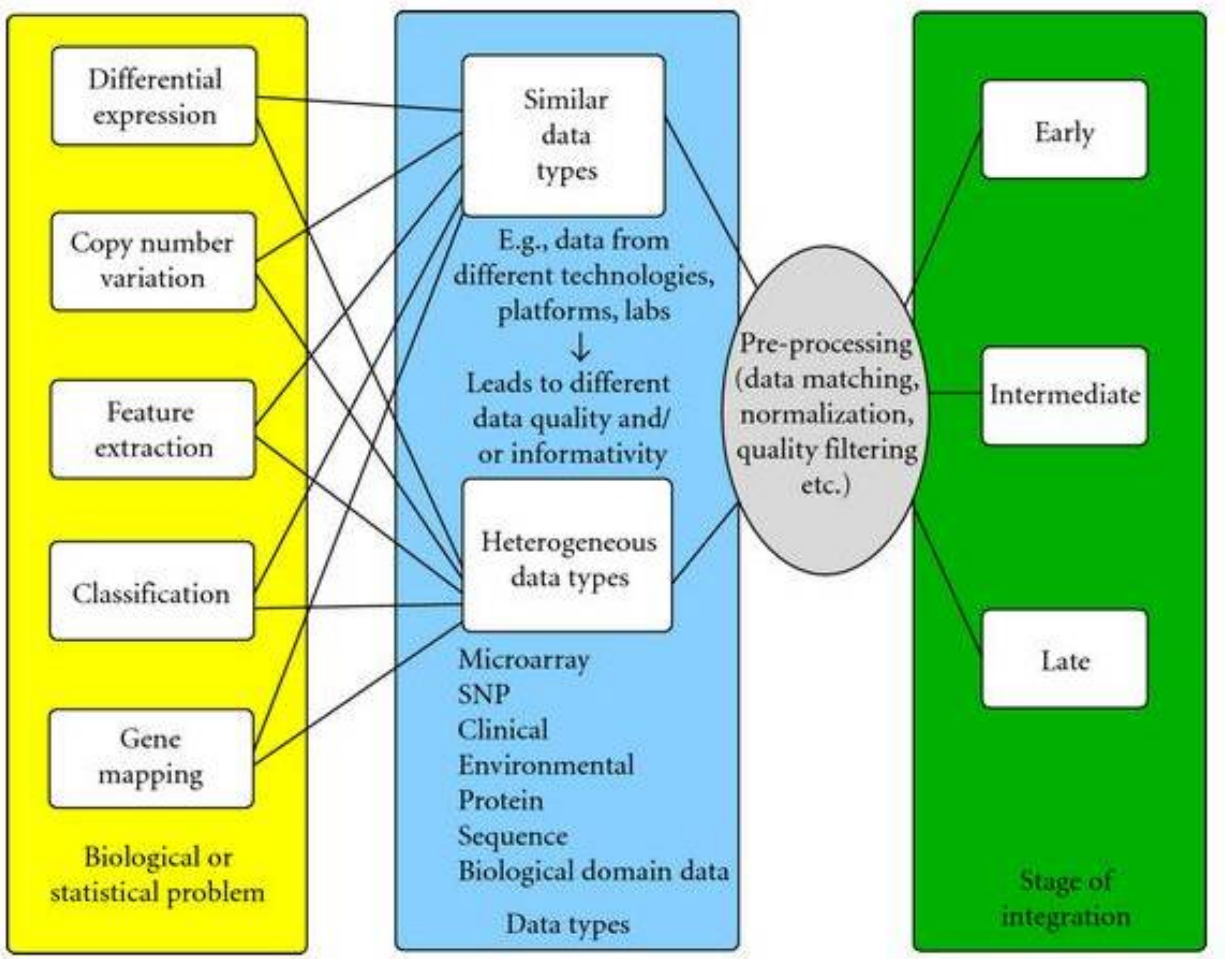
- "*Data integration*" may mean different things...
  - Computational combination of data
  - Combination of studies performed independently
  - Simultaneous analysis of multiple variables on multiple datasets.
  - Not to mention any possible approach for homogeneously querying heterogeneous data sources
- **Integrative analysis** may be preferable

Methods and Tools for IODA

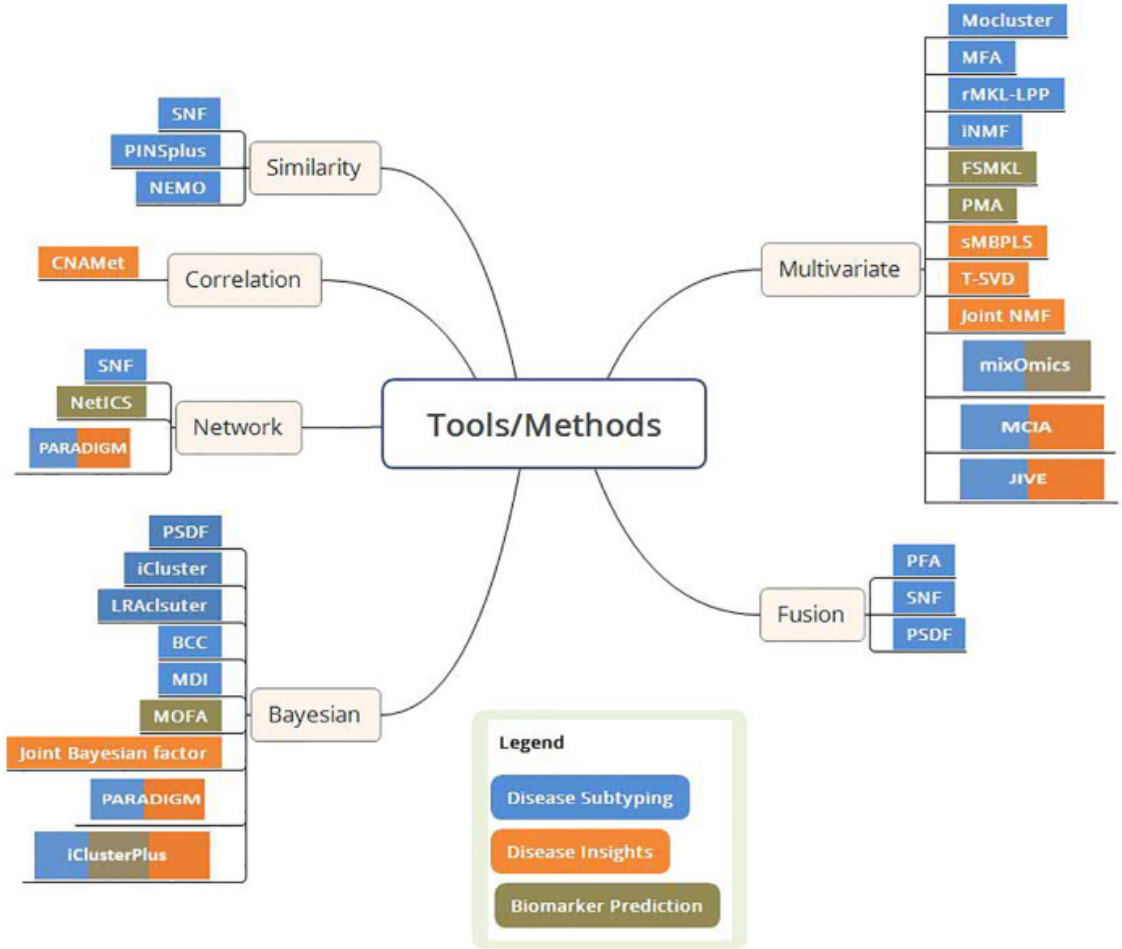
Integrative Omics Data Analysis

- The idea that efficient integration of data from different OMICS can greatly facilitate the discovery of true causes and states of disease is rapidly pervading the biomedical community
- The aims of integrative analysis is the deciphering of complex biological relationships empowered by the combined use of distinct pieces of information that represent a, probably partial, view of the different levels at which these processes happen

There are many types of integrative analysis



There are many methods ...



Subramanian et alt., 2020. *Multi-omics Data Integration, Interpretation, and Its Application*

There are many data repositories

DATA REPOSITORY	WEB LINK	DISEASE	TYPES OF MULTI-OMICS DATA AVAILAB
The Cancer Genome Atlas (TCGA)	https://cancergenome.nih.gov/	Cancer	RNA-Seq, DNA-Seq, miRNA-Seq, SN CNV, DNA methylation, and RPPA
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	https://cptac-data-portal.georgetown.edu/cptacPublic/	Cancer	Proteomics data corresponding to TC cohorts
International Cancer Genomics Consortium (ICGC)	https://icgc.org/	Cancer	Whole genome sequencing, genomic variations data (somatic and germline mutation)
Cancer Cell Line Encyclopedia (CCLE)	https://portals.broadinstitute.org/ccle	Cancer cell line	Gene expression, copy number, and sequencing data; pharmacological profiles of 24 anticancer drugs
Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)	http://molonc.bccrc.ca/aparicio-lab/research/metabric/	Breast cancer	Clinical traits, gene expression, SNP, CNV
TARGET	https://ocg.cancer.gov/programs/target	Pediatric cancers	Gene expression, miRNA expression, copy number, and sequencing data
Omics Discovery Index	https://www.omicsdi.org	Consolidated data sets from 11 repositories in a uniform framework	Genomics, transcriptomics, proteomic and metabolomics

Subramanian et alt., 2020. *Multi-omics Data Integration, Interpretation, and Its Application*

And many Data Visualization Portals

PORTAL NAME	OMICS DATA SUPPORTED	SOURCE REPOSITORY	ANALYSIS OF PRIVATE DATA	AVAILABILITY	REFERENCE
cBioPortal	Mutation, copy number, gene expression, miRNA expression, DNA methylation, protein abundance, and clinical data	TCGA and published studies (http://www.cbioportal.org/)	Yes	http://www.cbioportal.org/	Cerami et al[82]; Gao et al[83]
Firebrowse	Mutation, copy number, gene expression, miRNA expression, DNA methylation, protein abundance, and clinical data	TCGA	No	http://firebrowse.org/	NA
UCSC Xena	Copy number, somatic mutation, DNA methylation, gene and exon expression, protein expression, tissue specific expression data, PARADIGM pathway inference, and phenotype data	TCGA, CCLE, ICGC, GTEX, TARGET, and published studies	Yes	https://xena.ucsc.edu/	Goldman et al[84,85]
LinkedOmics	Clinical data, Copy number, miRNA expression, mutation, DNA methylation, gene expression, protein expression and abundance, phosphoproteome and glyco-proteome data	TCGA and CPTAC	No	http://www.linkedomics.org/	Vasaikar et al[86]
3Omics	Gene expression, protein and metabolite abundance	User data driven	Yes	https://3omics.cmdm.tw/	Kuo et al[87]
NetGestalt	Gene expression, mutation, and copy number data	TCGA, CPTAC, and published studies	Yes	http://www.netgestalt.org/index.html	Shi et al[88]
OASIS	Mutation, copy number, and gene expression data	TCGA, CCLE, GTEx, and published studies	No	http://www.oasis-genomics.org/	Fernandez-Banet et al[89]
Paintomics 3	Gene expression, miRNA expression, metabolite and region-specific ChIP-Seq, and Methyl-Seq data	User data driven	Yes	http://www.paintomics.org/	Hernández-de-Diego et al[90]
MethHC	DNA methylation, gene expression, and miRNA expression	TCGA	No	http://methhc.mbc.nctu.edu.tw/php/index.php	Huang et al[91]

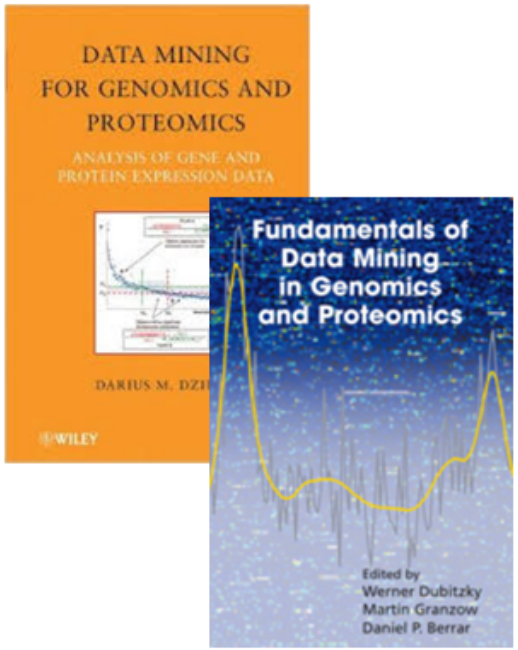
So what?

- We willl restrict to arbitrarily chosen situation
  - Multivariate statistical methods, classic and extensions
- for which we will sketch basic ideas,
  - General concept
  - Basic formulation
- and provide some examples of use

Multivariate statistics in genomics

- Multivariate methods have pervaded the field of genomics since its very beginning
  - The (in)fammous clustering (HC, heatmaps)
  - Matriz factorizations / Dimension reduction (PCA, SVD, CoA)
  - Discriminant Analyses (LDA $->$ DLDA, ...)

To cite but a few.



Best approach for omics data analysis?

- Classical Statistics
  - Multiple regression
  - Discriminant analysis
  - ANOVA
- Data tables are long and lean

- Assumptions
  - Independent variables
  - More observations than variables
  - Multivariate normality
  - Interested in one dependent
  - Few missings
- DO NOT hold for many omics data

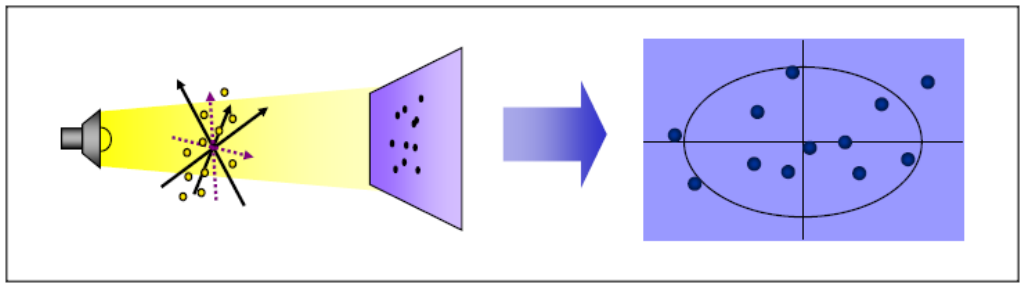
The nature of omics data

- Omics data are diverse
  - They measure distinct characteristics
    - GC/MS spectrum, Expression, Concentration...
- Although they have aspects in common
  - Most of them are high throughput
    - Many variables (K) measured simultaneously
  - Relatively expensive, ethical limitations, regulations
    - Few samples (N) analyzed

Figure: $K >> N$

A Better Way

- Multivariate analysis by projection (dimension-reduction, matrix decomposition) methods
  - Looks at ALL the variables together
  - Avoids loss of information
  - Finds underlying trends = "latent variables"
  - More stable models



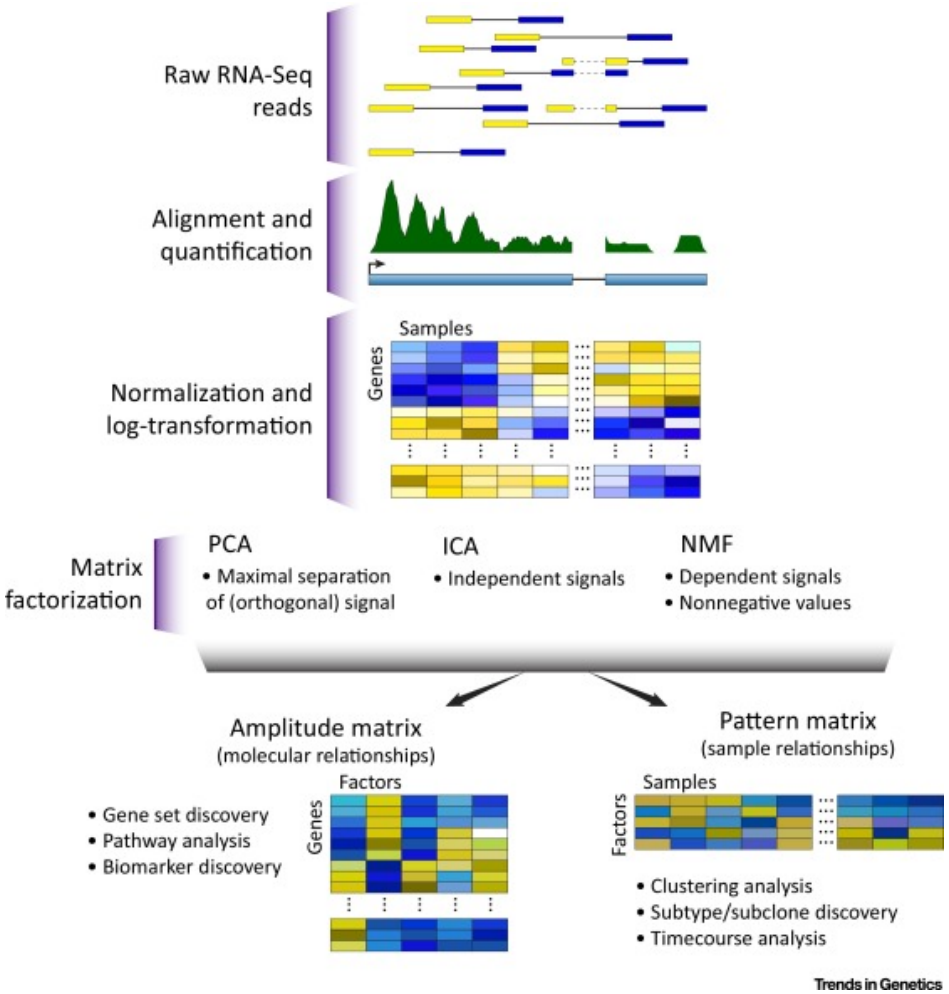
Principal Components Analysis

- PCA is an statistical method that can be traced back to Pearson (1901) or Hotelling (1933).
- Its main goals are
  - Capture the information provided by the original variables using a smaller number of new variables
  - Provide a representation of the data in reduced dimension
- PCA is intimately related with the most famous `Matrix Factorization` method.

What does PCA do?

- Given a $k \times n$ data matrix containing k (probably correlated) measurements on n samples (objects/individuals...), PCA decomposes this matrix in new k components such that they
  - account for different sources of variability in the data,
  - are uncorrelated, that is each component accounts for a different source of variability,
  - have decreasing explanatory ability: each component explains more than the following
  - allow for a lower dimensional representation of the data in terms of scores on principal components.
  - get an overview of the dominant patterns and major trends in the data (visualize clusters, identify outliers)

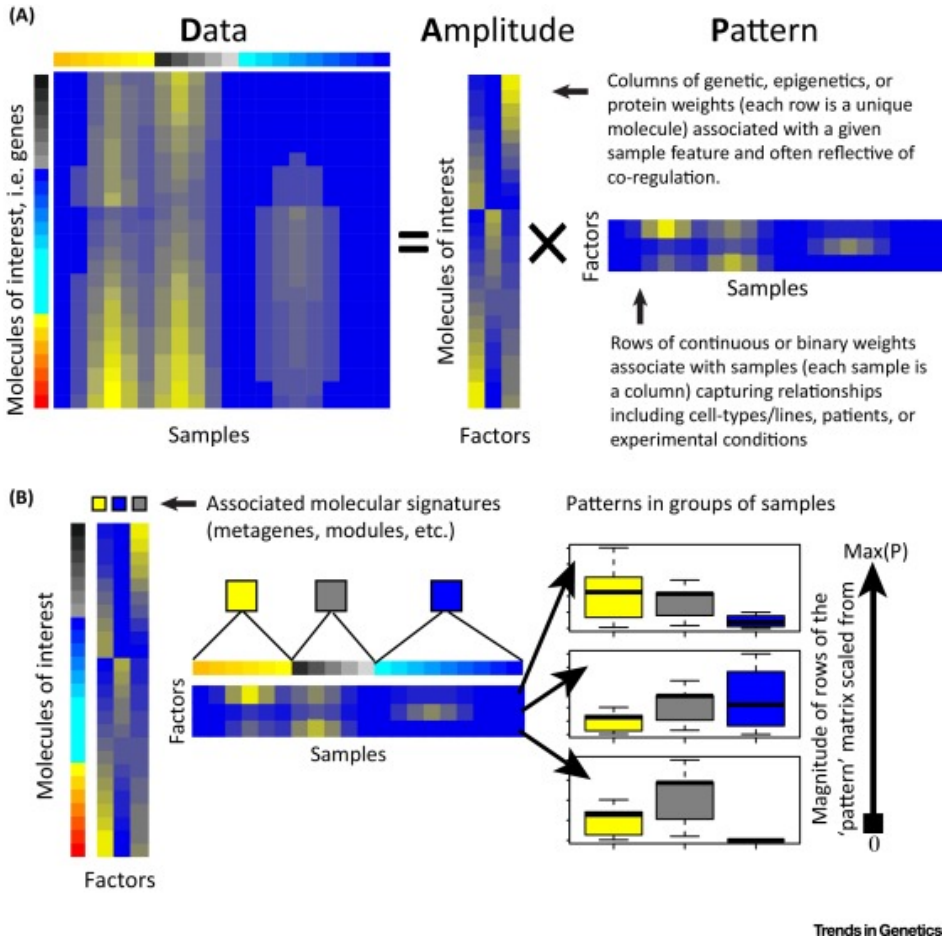
PCA is an example of *Matrix Factorization*



Stein-O'Brian et alt., 2017. *Enter the Matrix: Factorization Uncovers Knowledge from Omics*

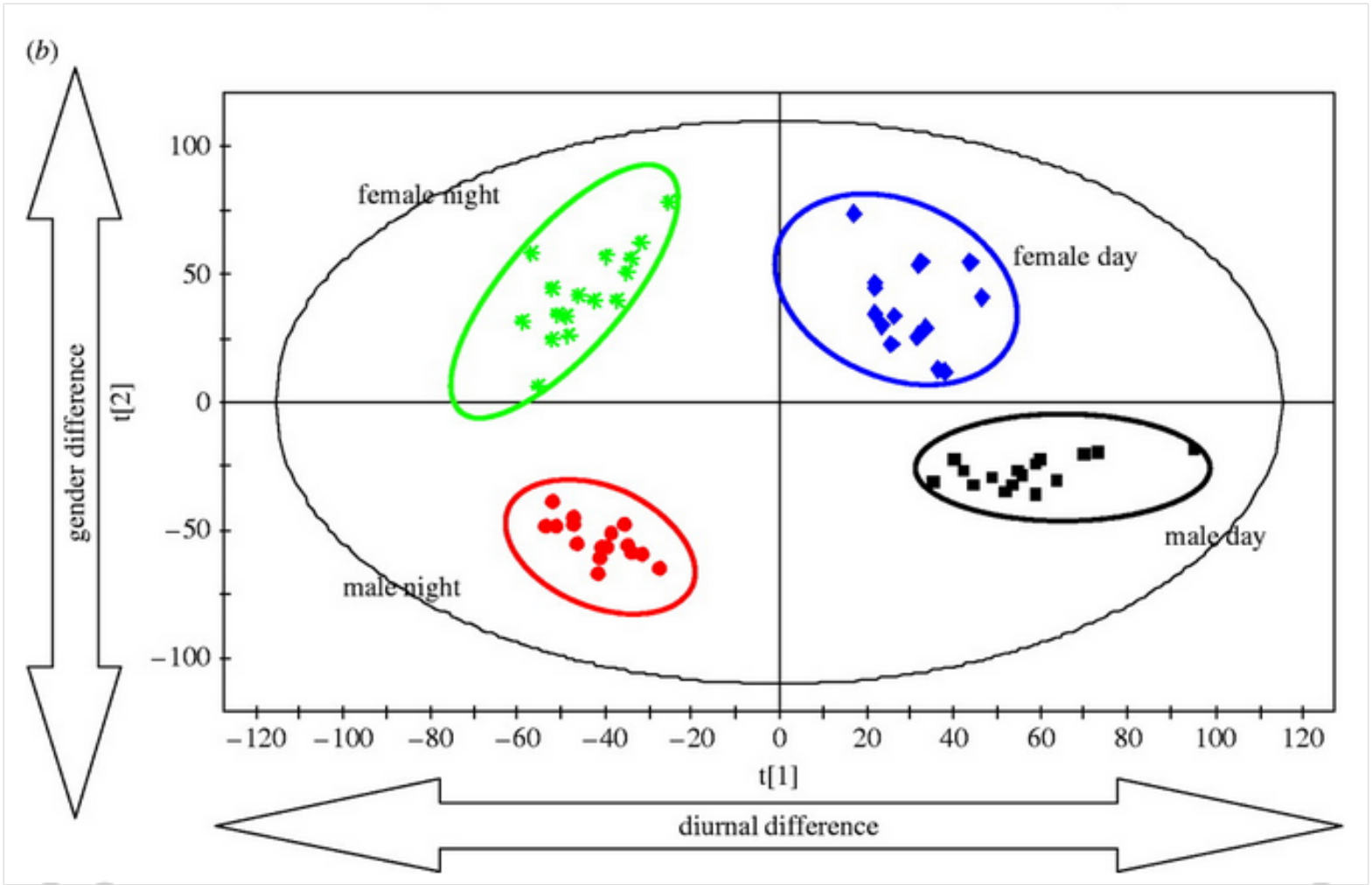
Matrix Product of Amplitude × Pattern Matrices

An Approximation the Preprocessed Input Data Matrix



Stein-O'Brian et alt., 2017. *Enter the Matrix: Factorization Uncovers Knowledge from Omics*

MFs provide data visualization in *reduced dimension*

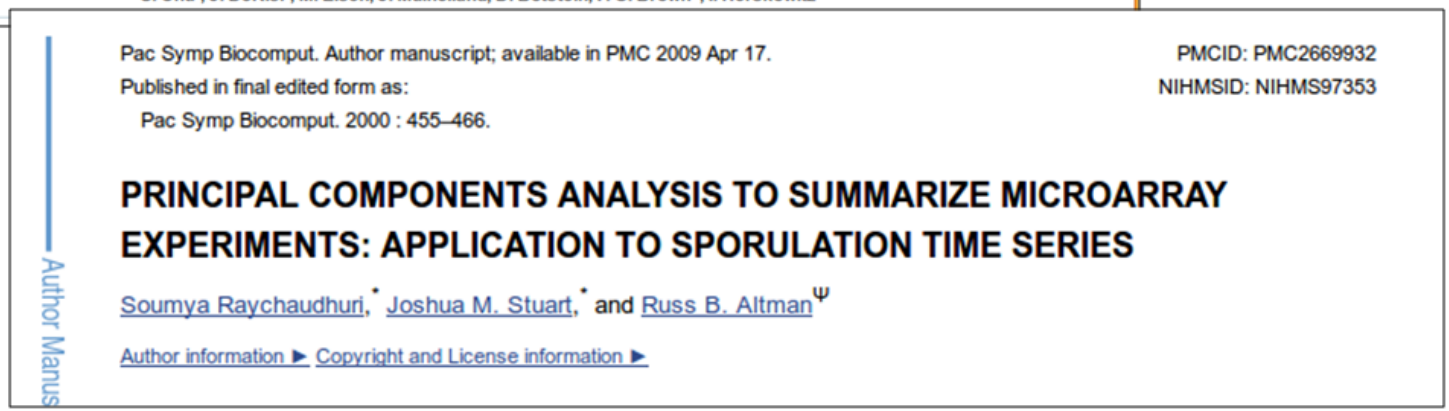


Applications of PCA

- PCA has been used in a wealth of applications
- The best known
  - Batch effect detection
  - Pattern recognition
- Besides standard applications there exist many extensions
  - Probabilistic PCA, Bayesian PCA, Inverse non-linear PCA, Nipals PCA, Robust PCA

Examples


Science
AAAS.ORG | FEEDBACK | HELP | LIBRARIANS
All Science Journals
Enter Search Term
CRAI. UNIVERSITAT DE BARCELONA ALERTS | ACCESS RIGHT
AAAS
NEWS
SCIENCE JOURNALS
CAREERS
MULTIMEDIA
COLLECTIONS
Science The World's Leading Journal of Original Scientific Research, Global News, and Commentary.
Science Home
Current Issue
Previous Issues
Science Express
Science Products
My Science
About the Journal
Home > Science Magazine > 23 October 1998 > Chu et al., 282 (5389): 699-705
Article Views
Abstract
Full Text
Full Text (PDF)
A correction has been published
Science 23 October 1998:
Vol. 282 no. 5389 pp. 699-705
DOI: 10.1126/science.282.5389.699
‹ Prev | Table of Contents | Next ›
RESEARCH ARTICLE
The Transcriptional Program of Sporulation in Budding Yeast
S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, I. Herskowitz


Pac Symp Biocomput. Author manuscript; available in PMC 2009 Apr 17.
Published in final edited form as:
Pac Symp Biocomput. 2000 : 455–466.
PMCID: PMC2669932
NIHMSID: NIHMS97353
PRINCIPAL COMPONENTS ANALYSIS TO SUMMARIZE MICROARRAY EXPERIMENTS: APPLICATION TO SPORULATION TIME SERIES
Soumya Raychaudhuri, Joshua M. Stuart, and Russ B. Altman
Author information ► Copyright and License information ►
Author Manus

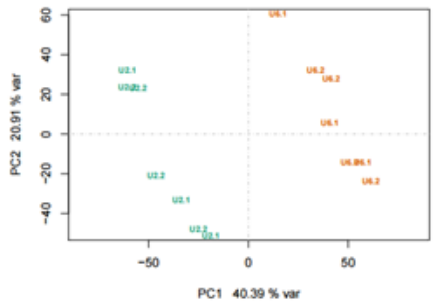
Examples

Journal of Proteomics
Volume 75, Issue 13, 16 July 2012, Pages 3938–3951

Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics

Josep Gregori[a, b], Laura Villarreal[a], Olga Méndez[a], Alex Sánchez[b, c], José Baselga[a], Josep Villanueva[a],

Show more



R Packages alpha Home Blog All packages Search R packages Searc

Exploratory Data Analysis of LC-MS/MS data by spectral counts

Exploratory data analysis to assess the quality of a set of LC-MS/MS experiments, and visualize de influence of the involved factors.

Author	Josep Gregori, Alex Sanchez, and Josep Villanueva
Date of publication	None
Maintainer	Josep Gregori <josep.gregori@gmail.com>
License	GPL-2
Version	1.2.0

View on Bioconductor

When single matrix factorization is not enough

- Ideally matrix factorizations can proide dimension-reduced data visualizations that help detect distinct patterns in features or samples.
- Sometimes the information in the data does not allow for this separation, but extending the factorization to **different omics** that can be *related differently* with the latent factors can do the job.
- As could be expected there exist many ways to do multiple factorization.
    - Multiple Factor Analysis, Regularized Generalized Canonical Correlation Analysis, Multiple Coinertia Analysis, ...

Example: Multi-omics analysis of colorectal cancer data

Integrative Analysis of Expression, Mutations and Copy Number Variation

- A set of 121 tumors from the TCGA (Weinstein, Collisson, Mills, et al. 2013) colorectal cancer cohort is analyzed.
- The tumors have been profiled for
  - gene expression using RNA-seq,
  - mutations using Exome-seq, and
  - copy number variations using genotyping arrays.
- Although two tumors arise in the colon, they may have distinct *molecular profiles*, which is important for treatment decisions.
- The subset of tumors used in this chapter belong to two distinct molecular subtypes CMS1, CMS3

Example: Multi-omics analysis of colorectal cancer data

Expression, Mutations and Copy Number Variation data

Example gene expression data (head)

	RNF113A	S100A13	AP3D1	ATP6V1G1
TCGA.A6.2672	21.19567	19.72600	11.53022	0.00000
TCGA.A6.3809	21.50866	18.65729	12.98830	14.12675
TCGA.A6.5661	20.08072	18.97034	10.83759	15.31325
TCGA.A6.5665	0.00000	11.88336	10.24248	19.79300
TCGA.A6.6653	0.00000	12.07753	0.00000	0.00000
TCGA.A6.6780	0.00000	12.99128	0.00000	19.96976

Example copy number data for CRC samples

	8p23.2	8p23.3	8p23.1	8p21.3
TCGA.A6.2672	0	0	0	0
TCGA.A6.3809	0	0	0	0
TCGA.A6.5661	0	0	0	0
TCGA.A6.5665	0	0	0	0
TCGA.A6.6653	0	0	0	0
TCGA.A6.6780	0	0	0	0

Example mutation data (head)

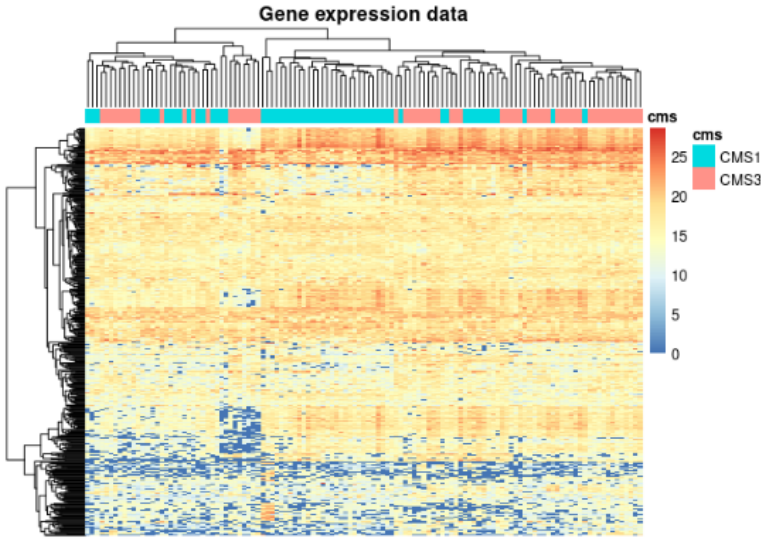
	TTN	TP53	APC	KRAS
TCGA.A6.2672	1	0	0	0
TCGA.A6.3809	1	0	0	0
TCGA.A6.5661	1	0	0	0
TCGA.A6.5665	1	0	0	0
TCGA.A6.6653	1	0	0	1
TCGA.A6.6780	1	0	0	0

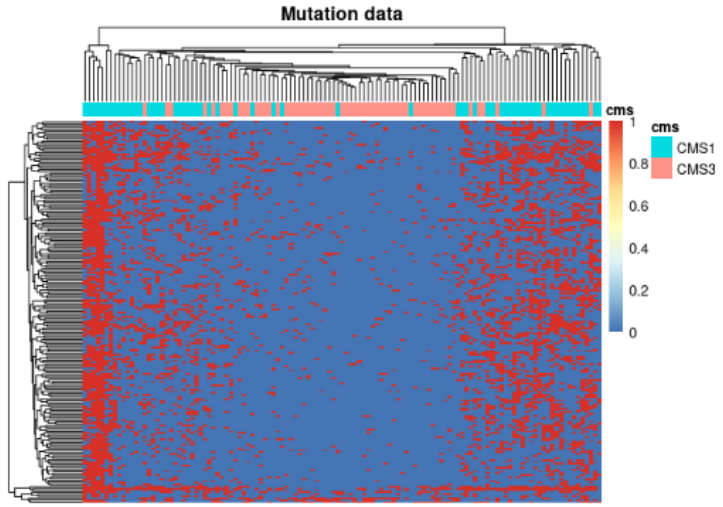
Clinical information (covariates)

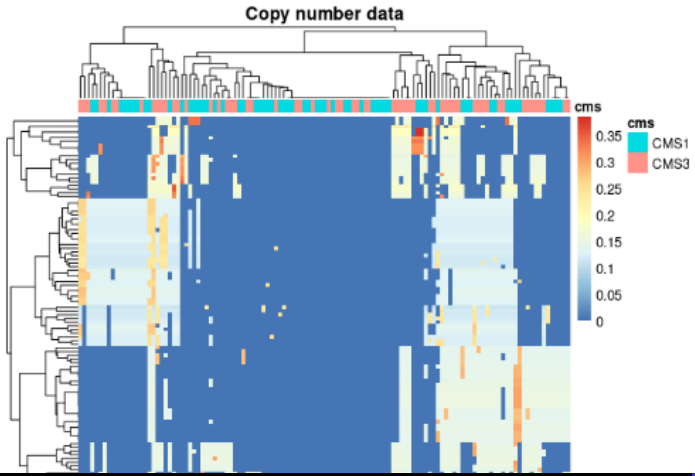
	cms
TCGA.A6.2672	CMS1
TCGA.A6.3809	CMS1
TCGA.A6.5661	CMS1
TCGA.A6.5665	CMS1
TCGA.A6.6653	CMS1
TCGA.A6.6780	CMS1

Example: Multi-omics analysis of colorectal cancer data

Individual omics do not separate well tumor types

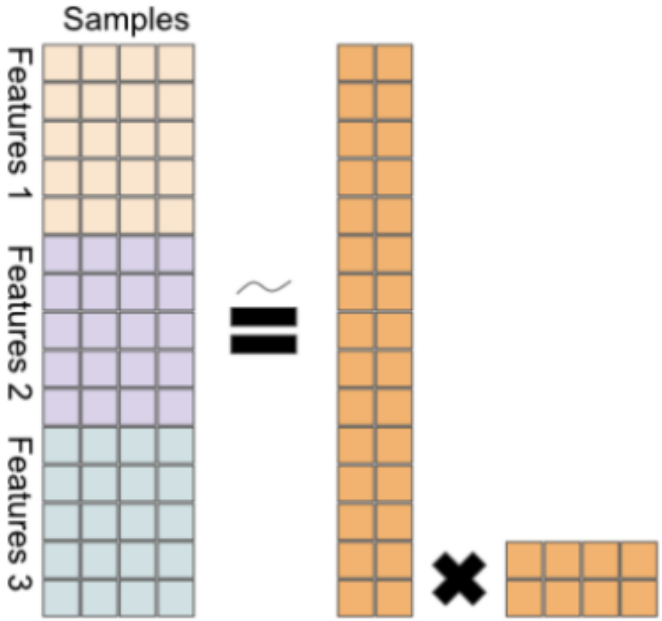


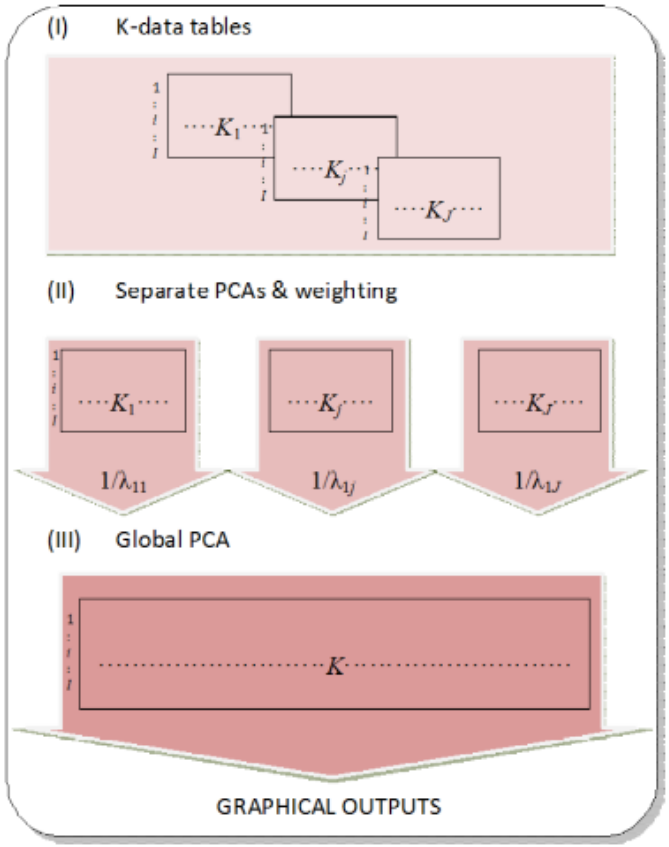




Multiple Factor Analysis

An extension of PCA to multiple data tables

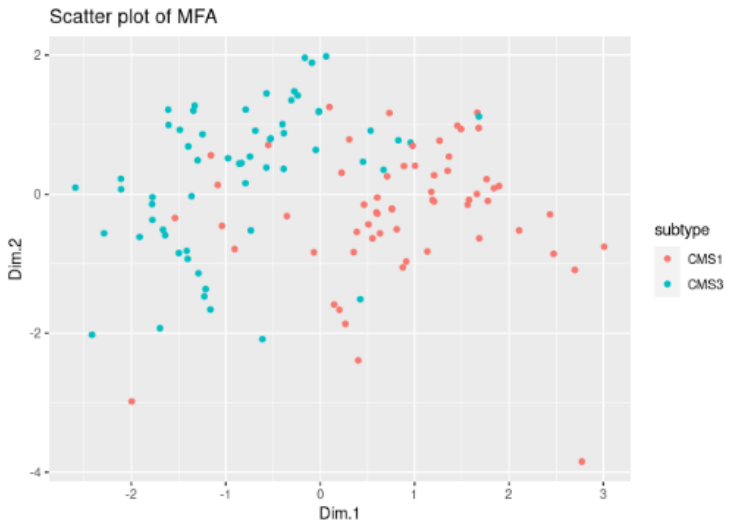


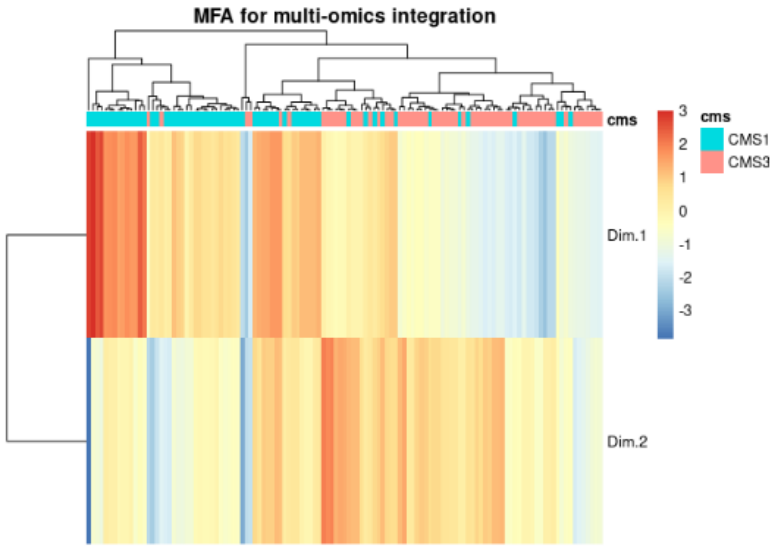


http://factominer.free.fr

Example: Multi-omics analysis of colorectal cancer data

Joint omics factorization provides much better separation





Are there alternatives?

There are many -or too many- alternative approaches

awesome-multi-omics

A community-maintained list of software packages for multi-omics data analysis.

While many of the packages here are marketed for "omics" data (transcriptomics, proteomics, etc.), other more general terms for this type of data analysis are:

- multi-modal
- multi-table
- multi-way

The common thread among the methods listed here is that the same samples are measured across different assays. The data can be described as multiple matrices/tables with the same number of samples and varying number of features.

The repo is in the style of Sean Davis' awesome-single-cell repo for single-cell analysis methods.

Contributions welcome...

For brevity, below lists only the first author of multi-omics methods.

Software packages and methods

Multi-omics correlation or factor analysis

- 2007 - **SCCA** - Parkhomenko - sparse CCA - paper 1, paper 2
- 2008 - **PCCA** - Waaijenborg - penalized CCA / CCA-EN - paper
- 2009 - PMA - Witten - Sparse Multi CCA - paper 1, paper 2
- 2009 - **sPLS** - Lê Cao - sparse PLS - paper
- 2009 - gesca - Hwang - RGSCA regularized generalized structured component analysis - paper
- 2010 - **Regularized dual CCA** - Soneson - paper
- 2011 - RGCCA - Tenenhaus - Regularized Generalized CCA and Sparse Generalized CCA - paper 1, paper 2
- 2011 - **SNMNMF** - Zhang - Sparse Network-regularized Multiple Non-negative Matrix Factorization - paper
- 2011 - scca - Lee - Sparse Canonical Covariance Analysis for High-throughput Data - paper
- 2012 - STATIS/DiSTATIS - Abdi - structuring three-way statistical tables - paper
- 2012 - **joint NMF** - Zhang - extension of NMF to multiple datasets - paper
- 2012 - **sMBPLS** - Li - sparse MultiBlock Partial Least Squares - paper
- 2012 - **Bayesian group factor analysis** - Virtanen - paper
- 2012 - RIMBANET - Zhu - Reconstructing Integrative Molecular Bayesian Networks - paper
- 2013 - FactoMineR - Abdi - MFA: multiple factor analysis - paper

https://github.com/mikelove/awesome-multi-omics

Summary I

- There is no universal "IODA" method
- Many families of many types of methods available: Need to be related, classified, filtered, benchmarked.
  - In many situations biology must guide the analysis
  - For example, miRNAs-mRNAs or mRNAS-Methylation
  - All data are not equally informative.
  - It is often the case that some omics dominate others
  - Gene expression and what else?

Summary II

- Promising approaches are those that
  - Allow inclusion of biological information,
  - Provide hints for interpretability
  - Implementations are available
  - E.g. See the MOFA package
- Biological question should be first.
- There are more mathematical/statistical tools than end-user bioinformatical solutions: Opportunity for developers
- Nothing said about ML/DL but clearly helpful in many situations.

Acknowledgements