Data Integration in Bioinformatics and Biomedicine

Alex Sanchez-Pla, Francesc Carmona, Ferran Reverter, Esteban Vegas, Pol Castellano
Genetics, Microbiology and Statistics Department (UB)
Statistics and Bioinformatics Unit (VHIR)
2021-07-08

Outline

1) Data, Data-related Projects and Data Science

2) Integrative Omics Data Analysis

3) Open problems we work in

4) Biomedical data integration and data sharing

5) Data Projects we work in

6) Looking ahead

Data, Data-related Projects and Data Science

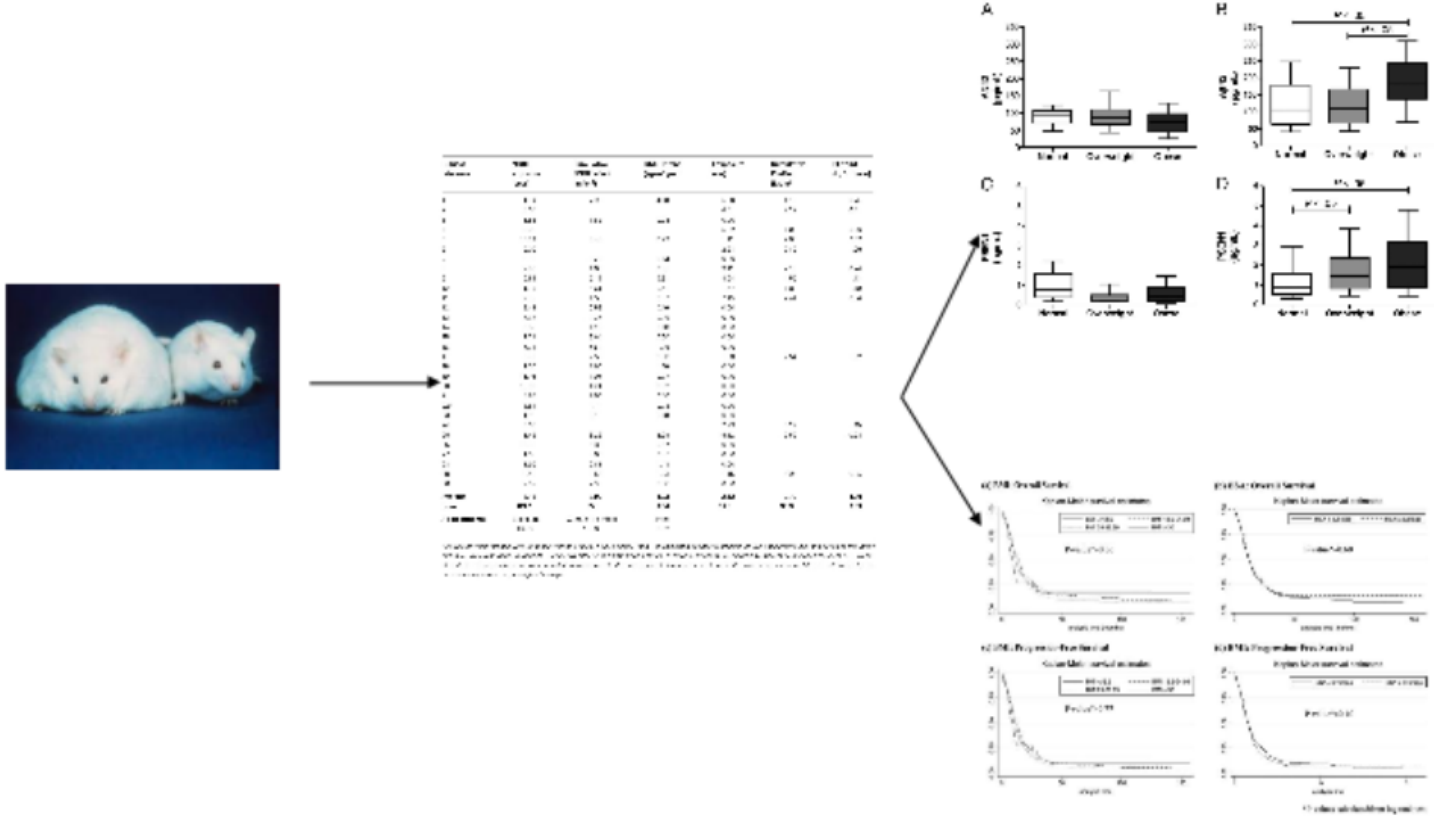
The importance of Data

Data is essential to solve problems or answer questions

- Simplistically, when we wish to solve a problem we collect data
  - Statistics: Questions --> [Hypothesis] --> Get data
  - Data Science: Questions --> Get Data
- Data availability and complexity increases all the time
  - More (Big) data --> Better solutions?
- 21st century holy grial:
  - **Data + AI --> Personalized medicine**

From XXth century studies

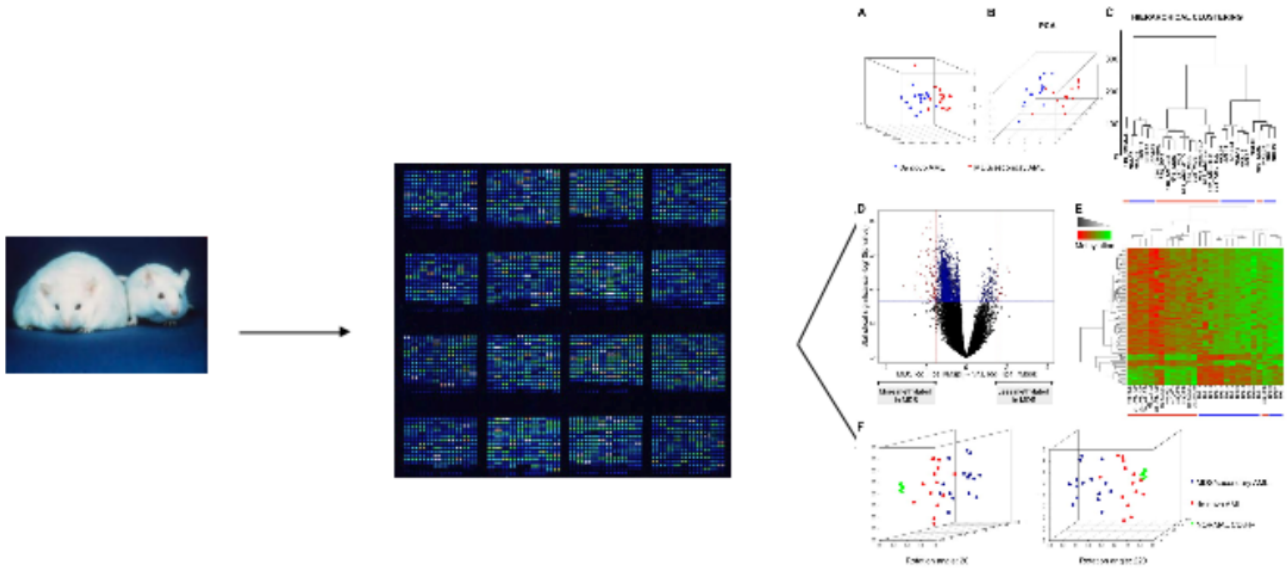
Traditional Statistical Analysis



To beginning of XXth

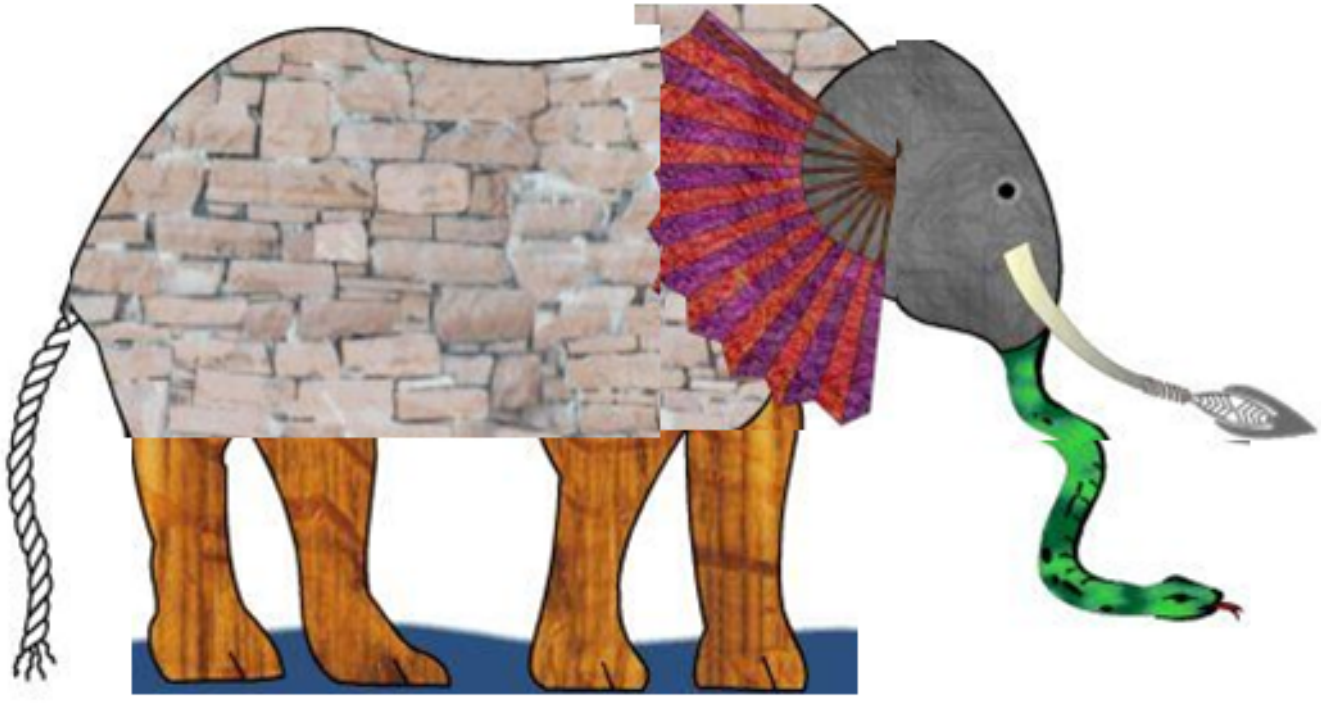
Microarrays: One dataset, *many* genes

Efron called the XXIst century "*The century of microarrays*"



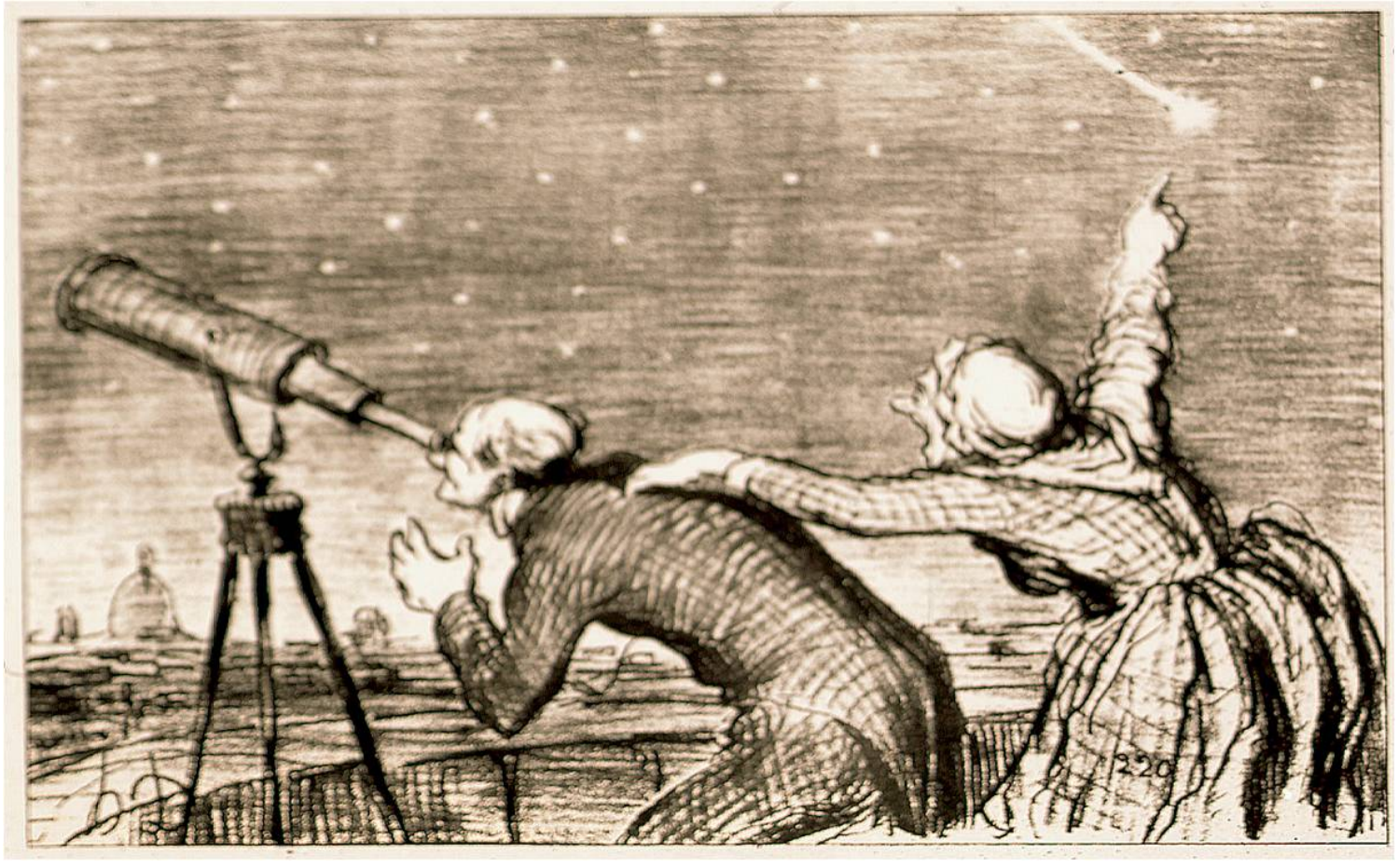
Why integration?

The whole is more than the sum of parts



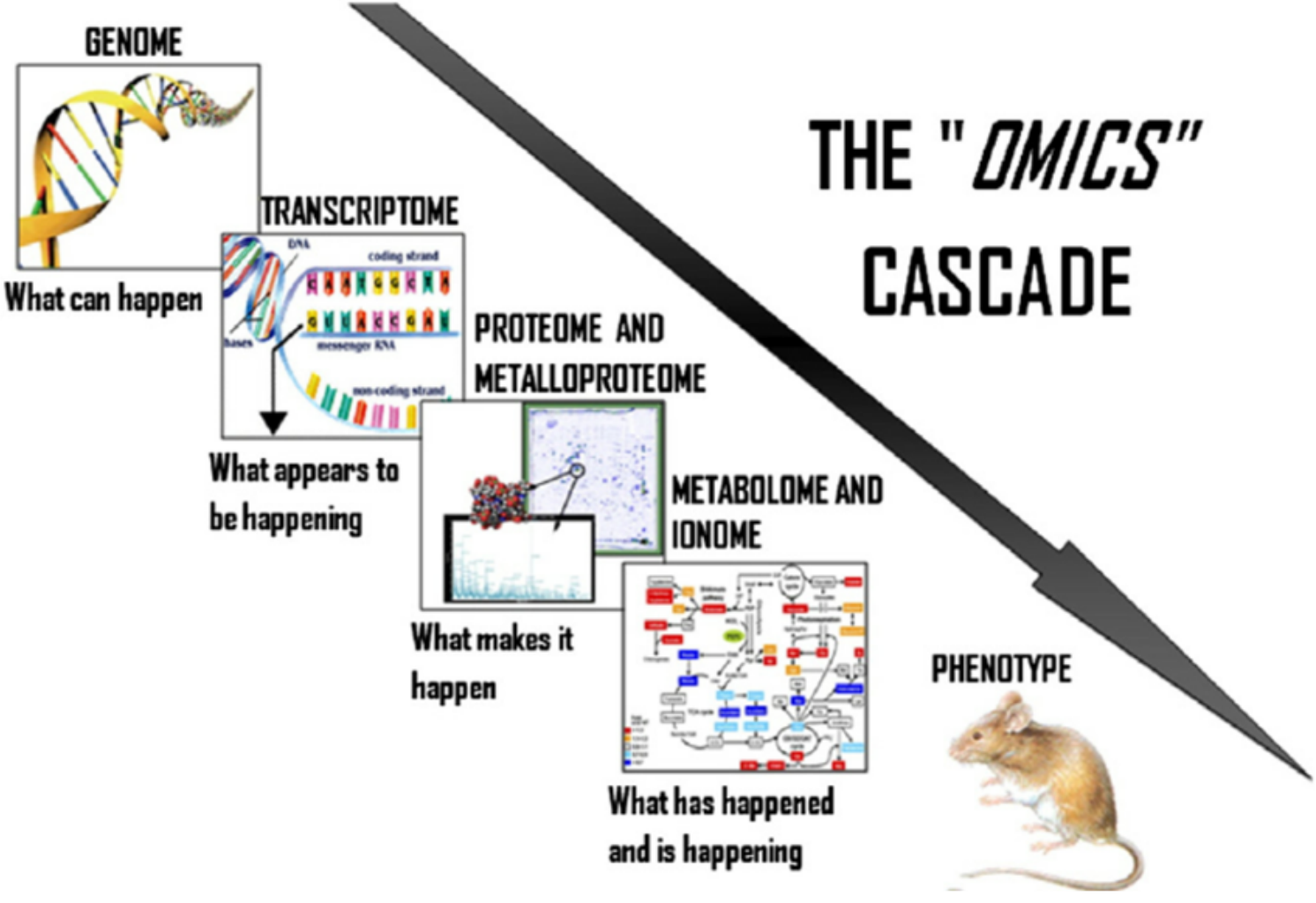
Why integration?

We May lose important information



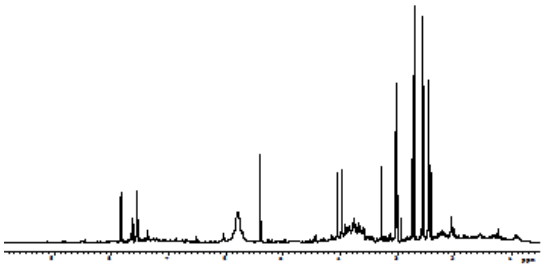
Why integration?

Don't forget the Omics Cascade

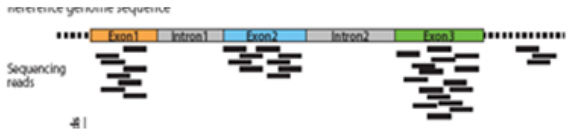


From single to multi-omics

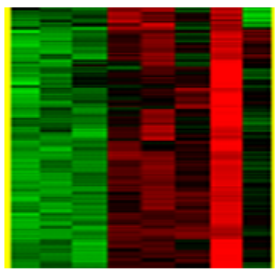
Integromics entered the game (*We want it all*)



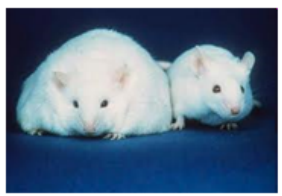
H NMR metabolites

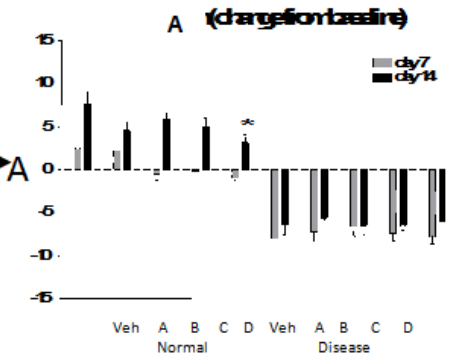


"NGS-Sequences

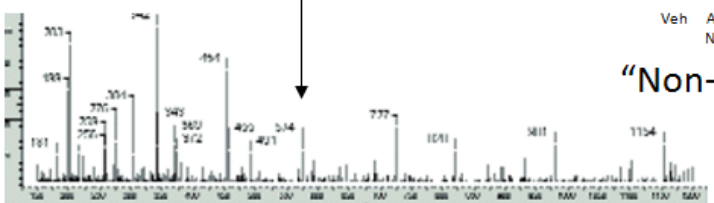


Affy Transcriptome





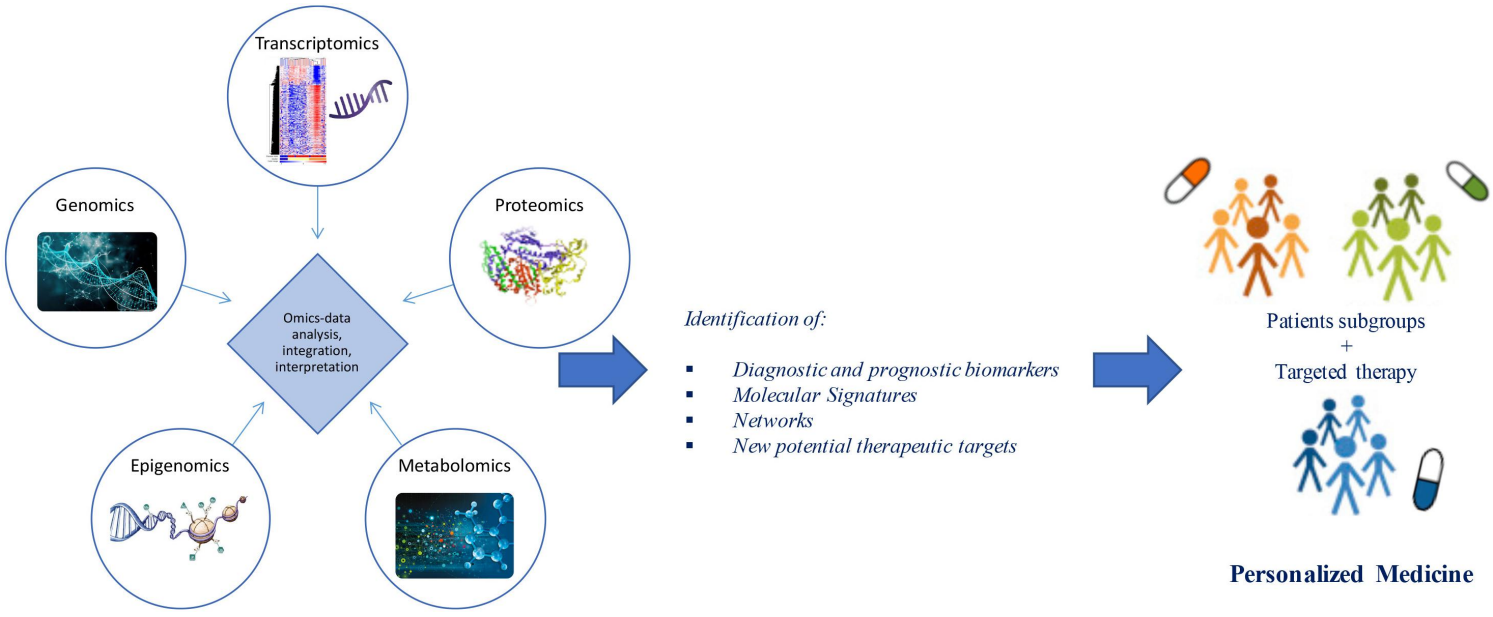
"Non-omic" markers



LC-MS proteomicss

Can we get more data?

From MultiOmics to Personalized Medicine

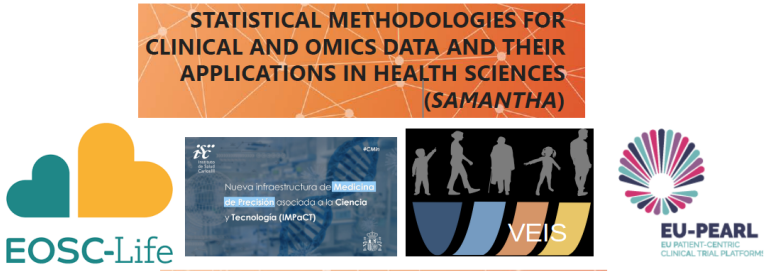


Google "multi-omics for personalized medicine" images to see more ...

Data-Centric Projects have become key

- The need to have multimodal data, including omics data, images, and clinical data has driven a multitude of **Research Projects**
- Their first or ultimate objective is to *facilitate the path to personalized medicine* through:
    - data integration
        - data sharing,
        - data exchange (OMOP, HL7)
        - data FAIRification,
        - cloud computing and
        - (federated) data analysis,

        among others.
- This represents an *unprecedented opportunity*.
- But requires the right people and knowledge.

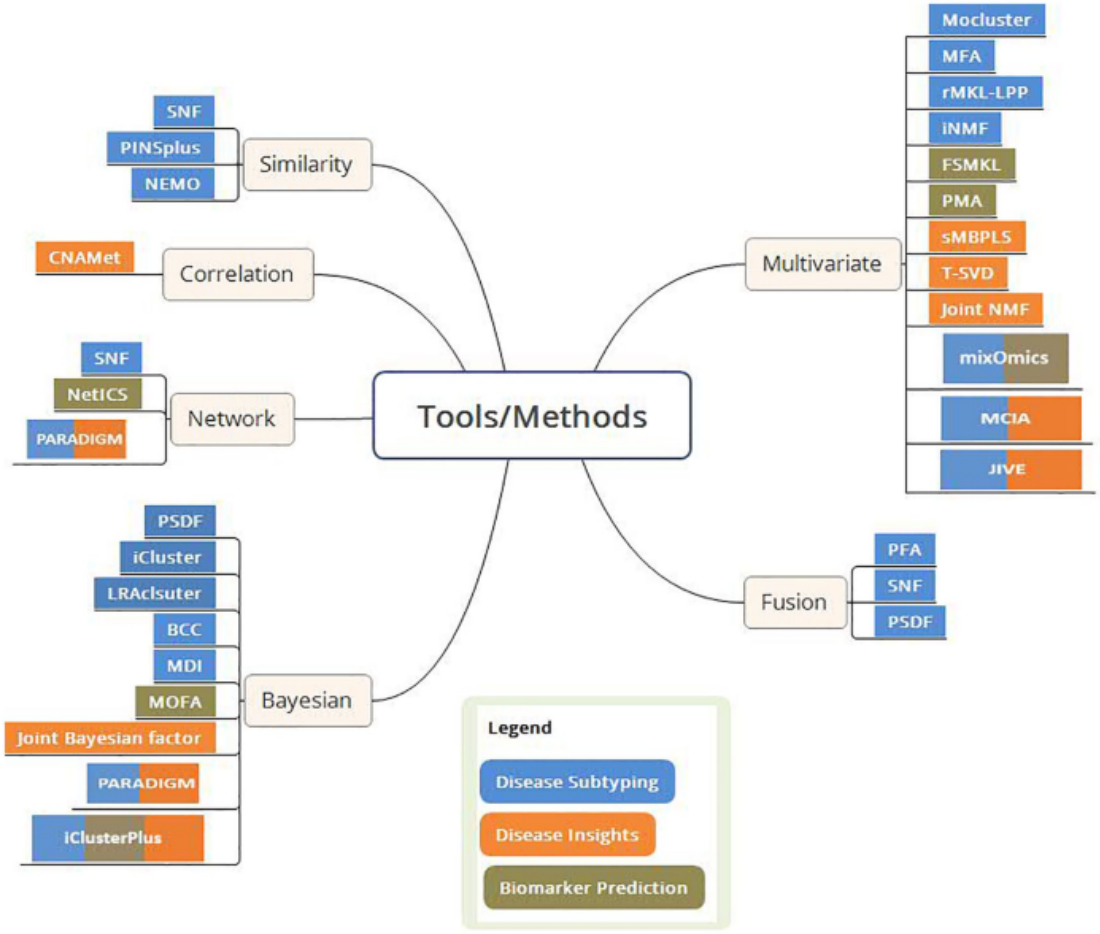


Integrative Omics Data Analysis

The goal(s) of Integrative Omics

- The idea that efficient integration of data from different OMICS can facilitate the discovery of true causes and states of disease has pervaded the biomedical community.
- The general gol of integrative analysis is the *deciphering of complex biological relationships (CBP)* 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*.
- More specifically integrative omics data anlysis is applied for
    - Disease subtyping,
    - Disease insights,
    - Biomarker development,
    - Combination with non-omics

Methods of integrative analysis



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

Many data repositories

DATA REPOSITORY	WEB LINK	DISEASE	TYPES OF MULTI-OMICS DATA AVAILABLE
The Cancer Genome Atlas (TCGA)	https://cancergenome.nih.gov/	Cancer	RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, and RPPA
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	https://cptac-data-portal.georgetown.edu/cptacPublic/	Cancer	Proteomics data corresponding to TCGA 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, and 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, proteomics, and metabolomics

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

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]

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

So what?

- We willl restrict to arbitrarily chosen situations:
  - Multivariate statistical methods, classic and extensions
- for which we will sketch,
  - General ideas
  - Use case
- and provide some examples of use.
  - See workshop

General framework: Matrix factorization

- Matrix factorizations have become very popular in fields such as machine learning, recommender systems or deep learniong.

A 2020 Vision of Linear Algebra

$A = CR =$ [ ] [ ] Independent columns in C

$A = LU =$ [$\searrow$ 0] [0 $\searrow$] Triangular matrices L and U

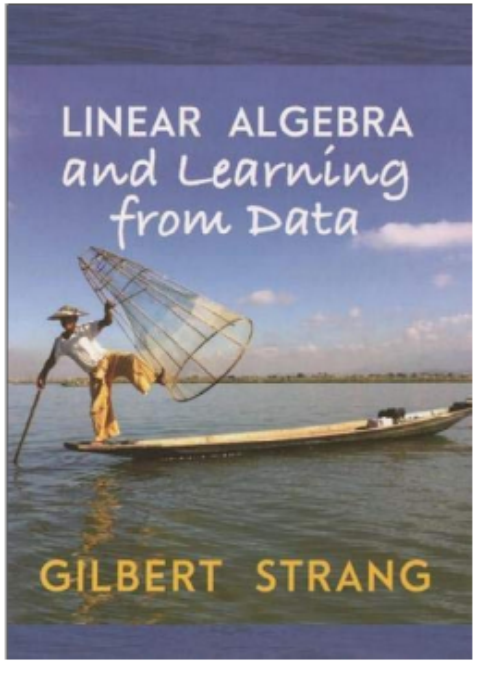
$A = QR =$ [q_1 q_n] [0 $\searrow$] Orthogonal columns in Q

$S = Q\Lambda Q^{\mathrm{T}}$ $Q^{\mathrm{T}} = Q^{-1}$ Orthogonal eigenvectors $Sq = \lambda q$

$A = X\Lambda X^{-1}$ Eigenvalues in Λ Eigenvectors in X $Ax = \lambda x$

$A = U\Sigma V^{\mathrm{T}}$ Diagonal Σ = Singular values $\sigma = \sqrt{\lambda(A^{\mathrm{T}}A)}$

Orthogonal vectors in $U^{\mathrm{T}}U = V^{\mathrm{T}}V = I$ $Av = \sigma u$

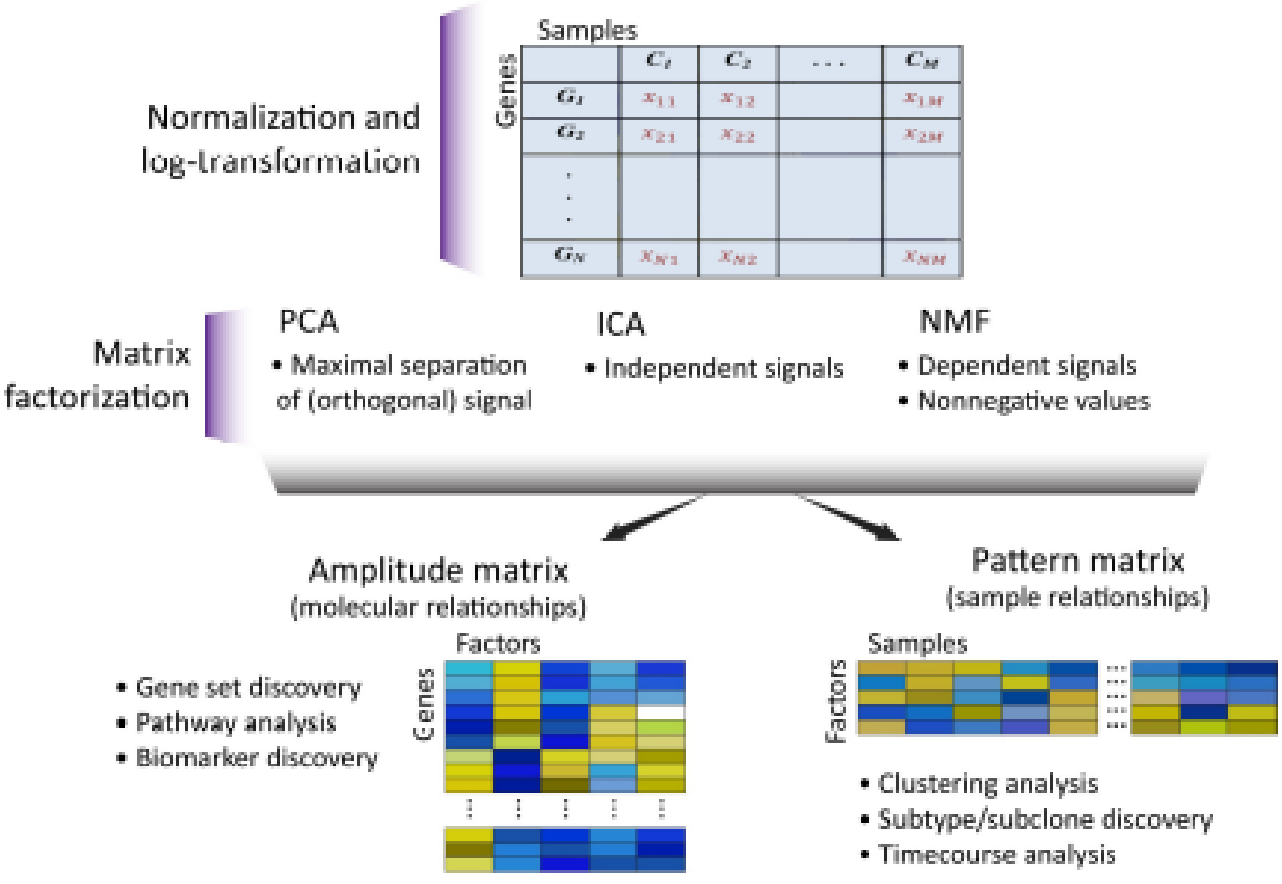


"Many key ideas of linear algebra, when you look at them closely, are really factorizations of a matrix, where the original matrix becomes the product of 2 or 3 special matrices."
Gilbert Strang.

Omics technolgies yield data matrices

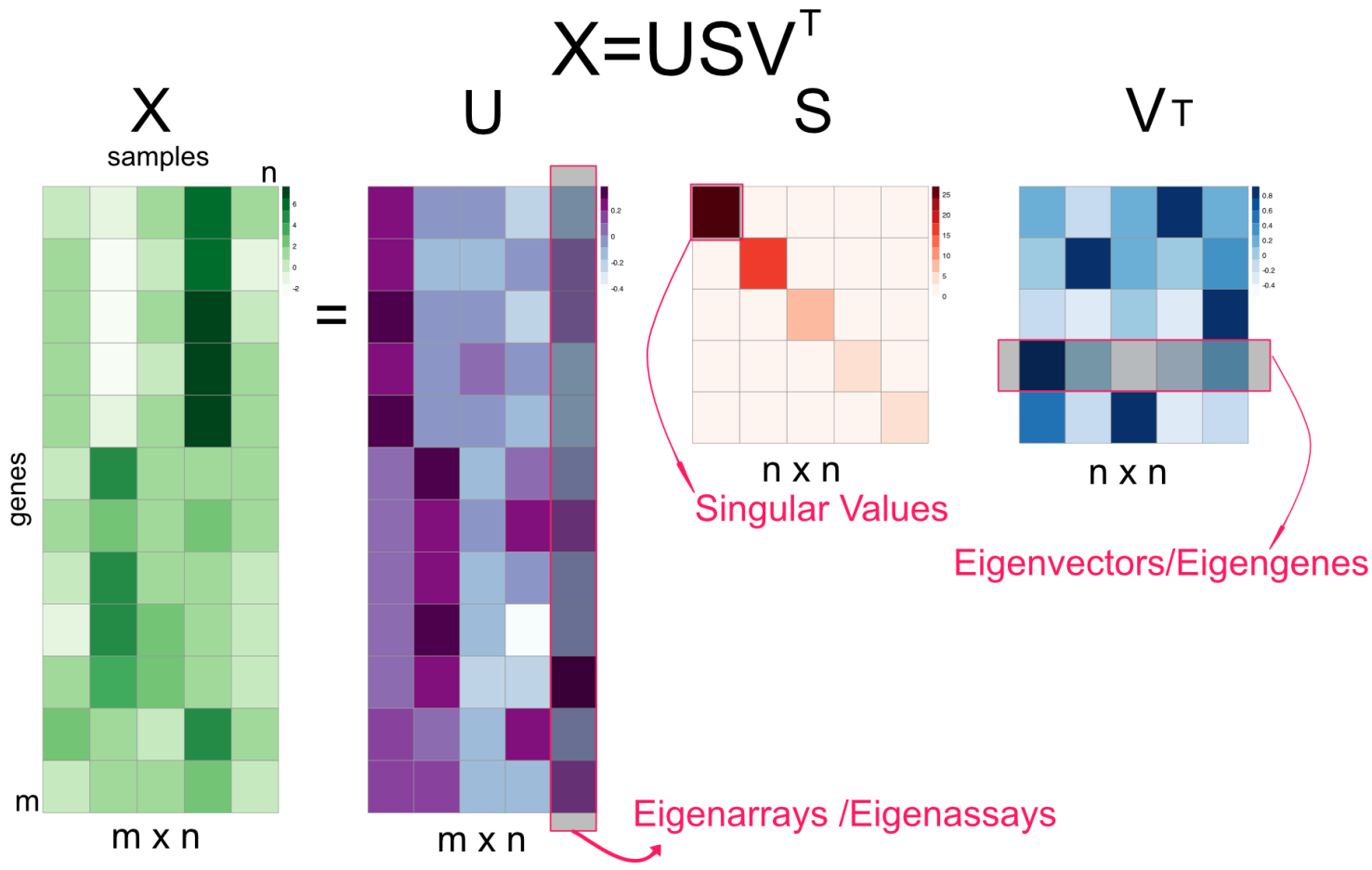
Matrices can be factored

- Factorization helps discover the "true" biological dimension
- by revealing hidden complex biological processes (CBPs)



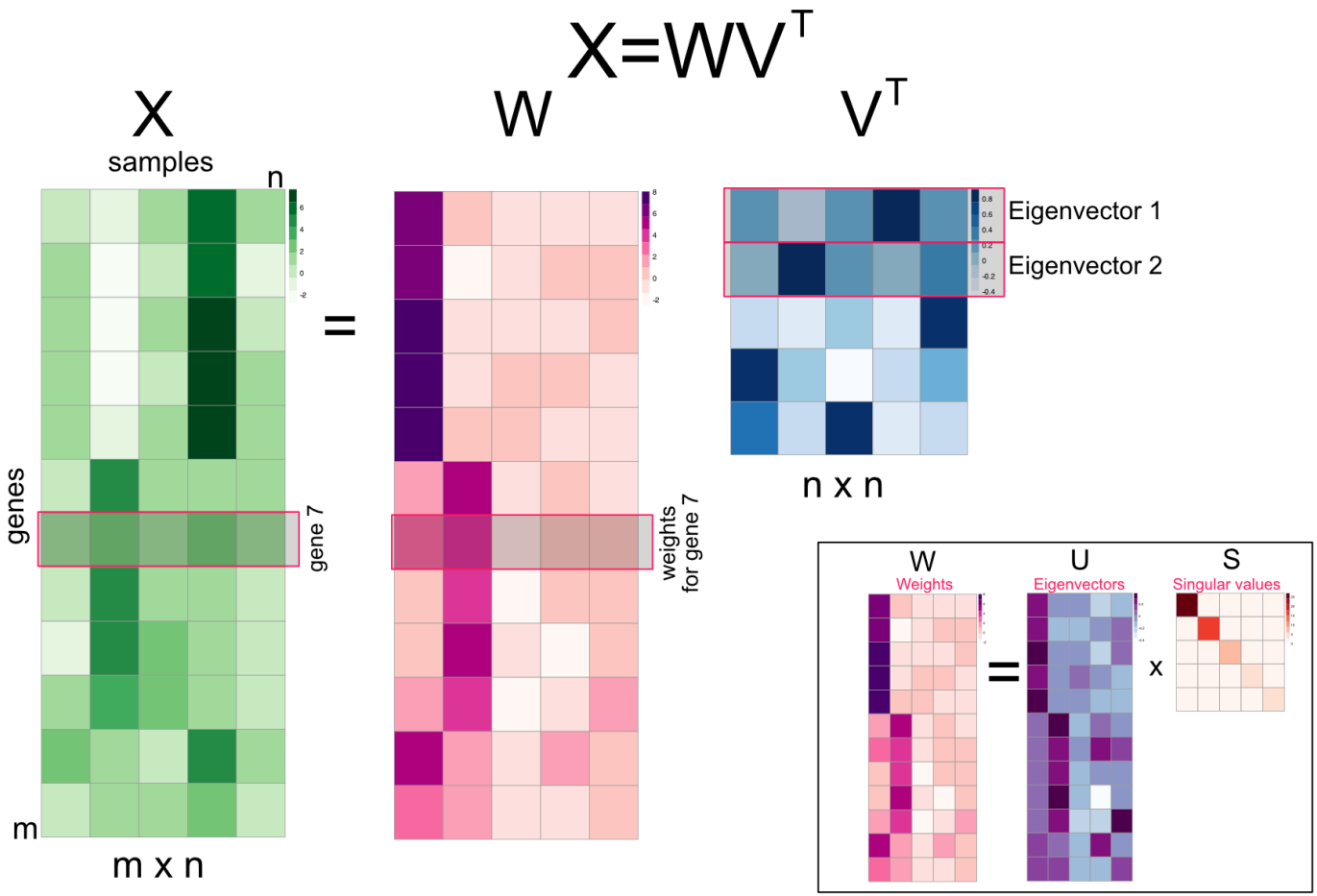
The SVD

The mother of all factorizations

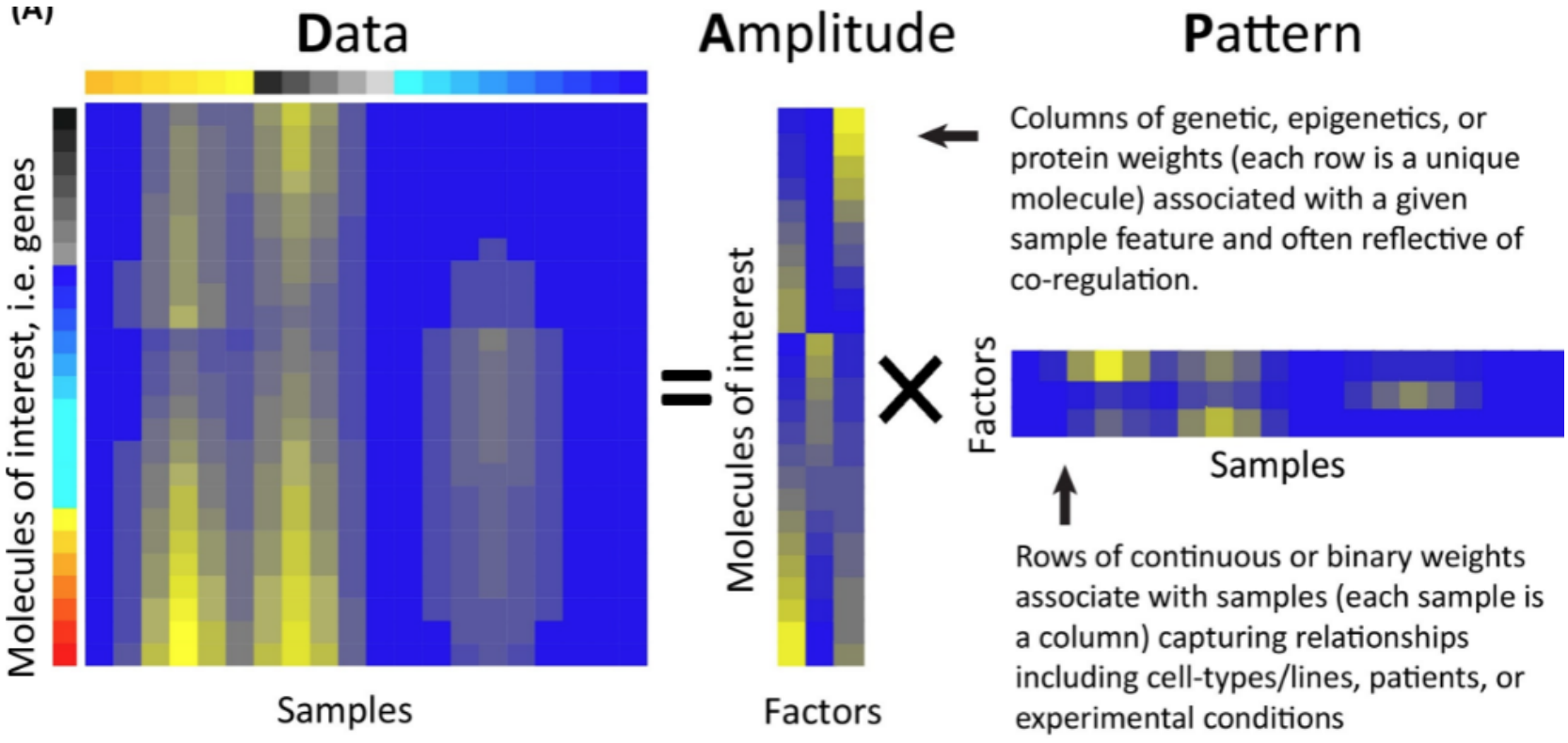


The SVD

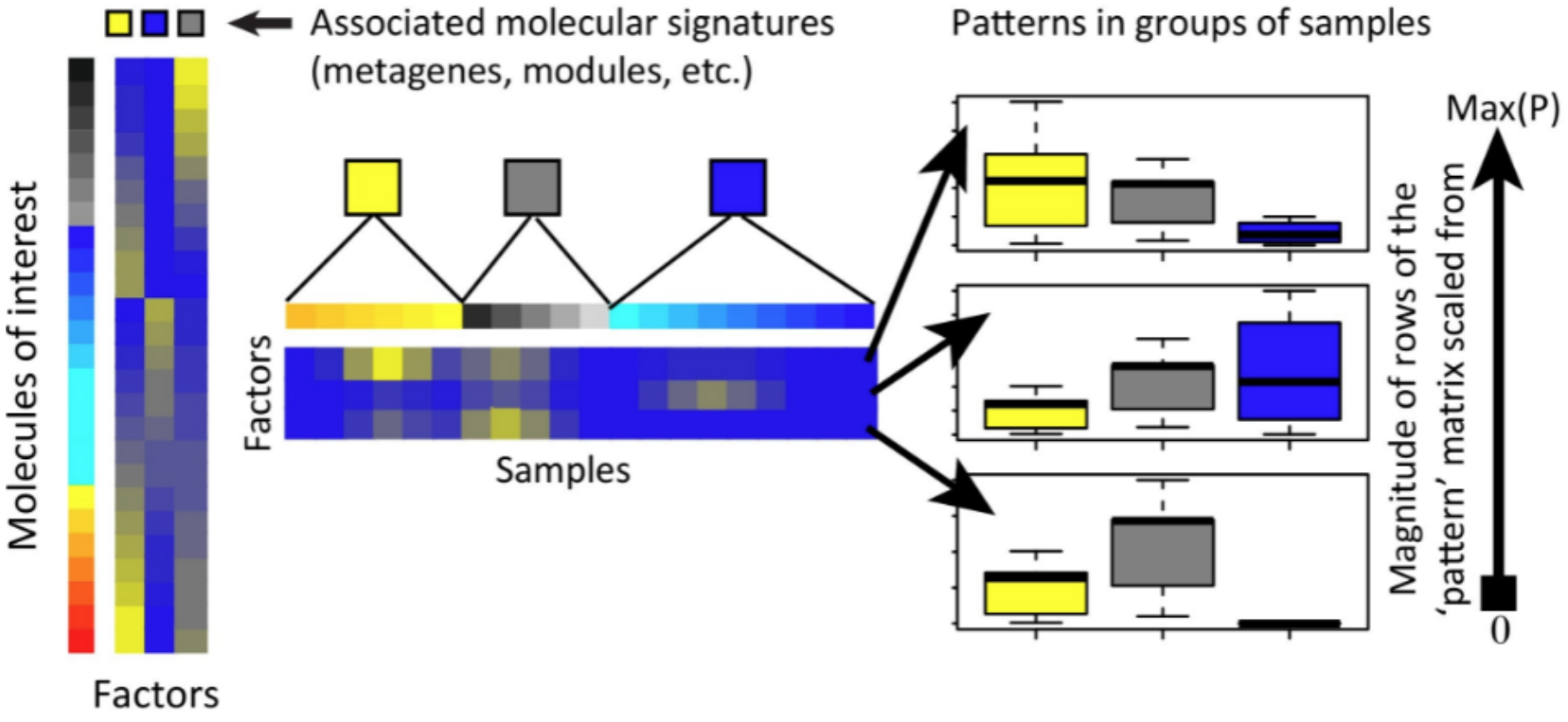
SVD can be re-written as latent factors



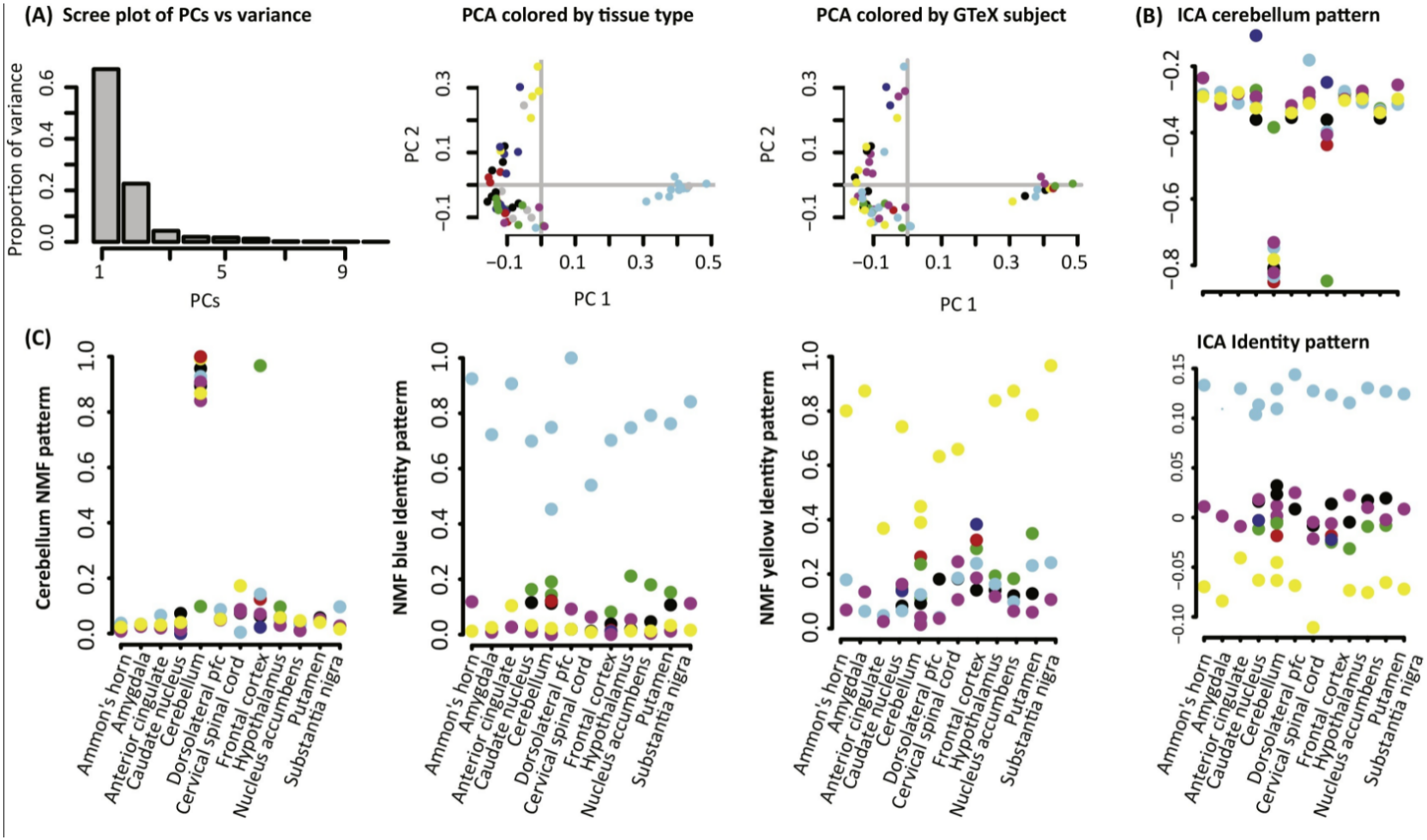
Factorization helps interpretation



Factorization helps interpretation



Distinct factorizations, distinct views

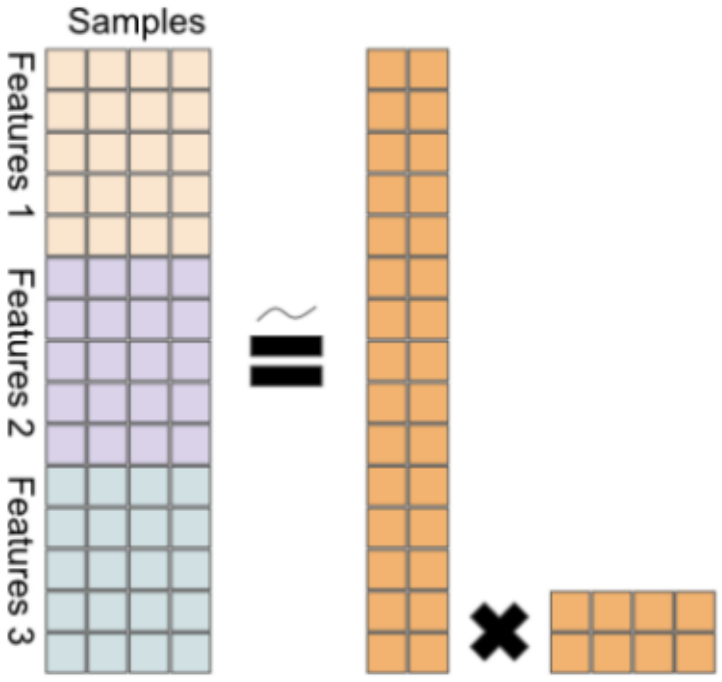


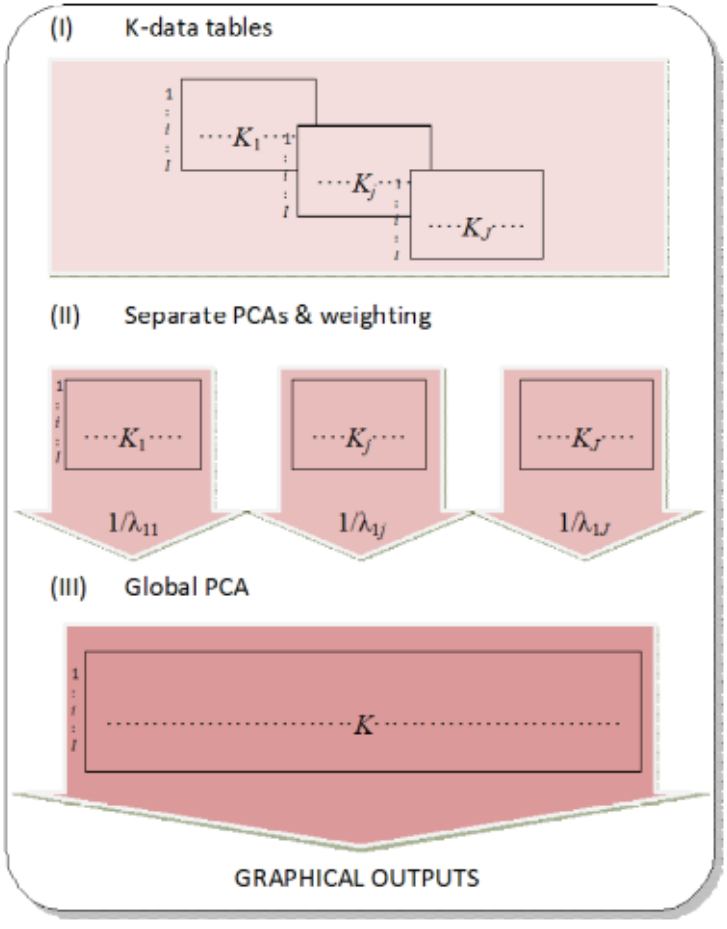
When a single omics is not enough

- Ideally matrix factorizations can provide 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, ...

Multiple factor Analysis

Straightforward generalization of PCA





A Colon Cancer Example

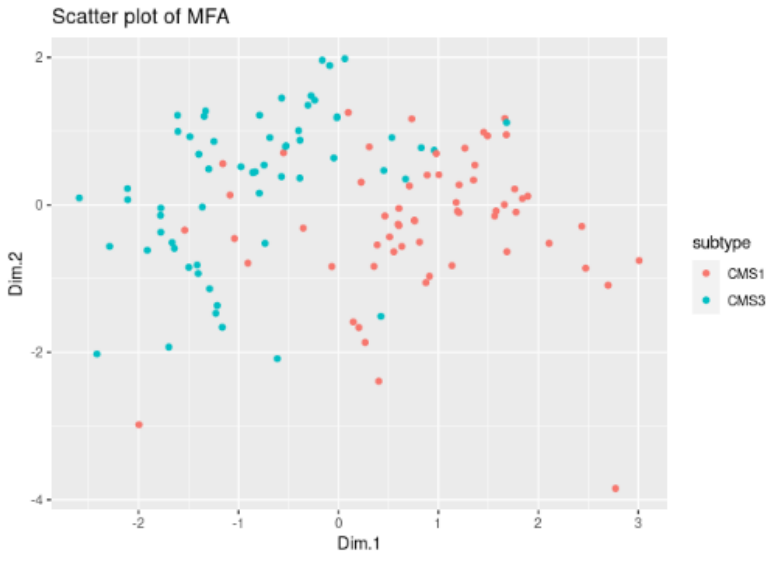
- A set of 121 tumors from the TCGA (Weinstein, Collisson, Mills, et al. 2013) colorectal cancer cohort is analyzed.
- The tumors have been profiled for
  - gene expression using RNA-seq,
  - mutations using Exome-seq, and
  - copy number variations using genotyping arrays.
- Although two tumors arise 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

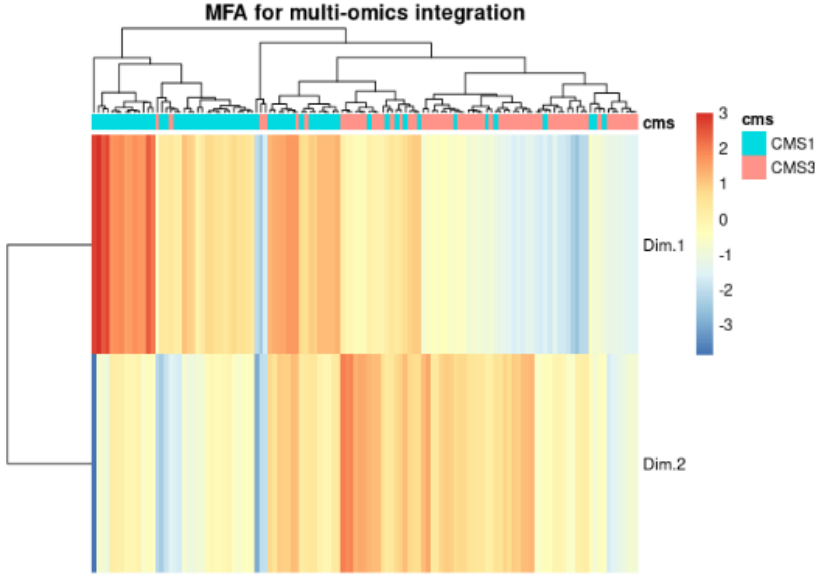
A Colon Cancer Example

Single omics don't separate well

A Colon Cancer Example

Joint omics factorization provides much better separation





Summary part I

- Biological Processes are complex and they are "probably" better revealed using distinct sources of information.
- While there are many approache for the integrative analysis 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
    - All data are not equally informative and it is often the case that some omics dominate others. : "Gene expression and *what else*?"

The workshop

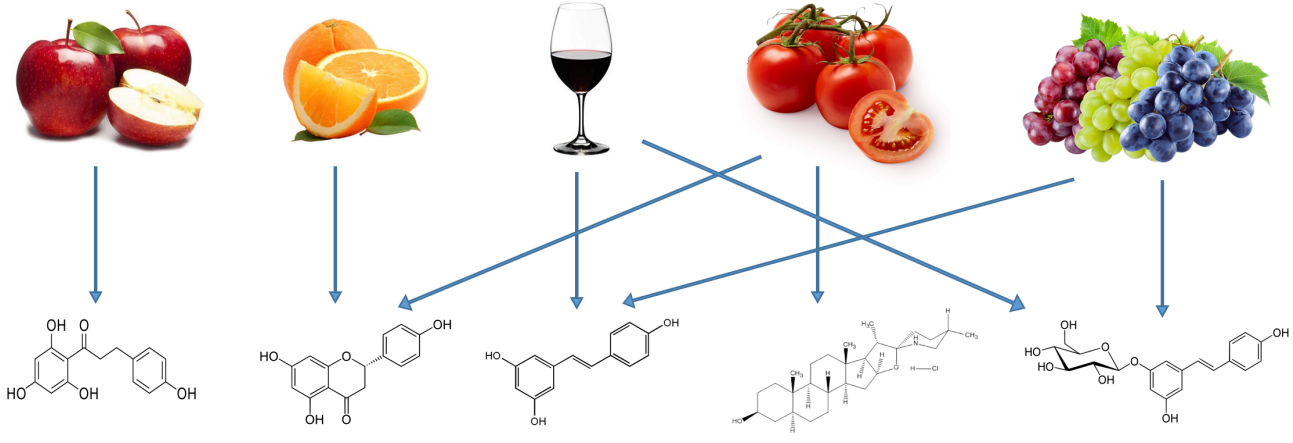
- Details and examples of single and multi-omics analyses are provided in the workshop.
- Access https://github.com/alexsanchezpla/202107-GRBio-Integration-Workshop for the workshop materials.
    - You can clone or download the repository (see previous GRBio workshops)
    - Run the `InstallPackages.R` script in your computer to be sure all packages can be installed
- **Warning** the materials will be updated. Be sure to refressh the repository tomorrow at 9:30

Integration problems we work with

FOBI (Food-Biomarker Ontology)

Pol Castellano, Alex Sánchez

- Heterogeneus nutritional data (semantic problem) -> **FoodOn**
- Difficult association of nutritional data with other types of data (semantic and quantitative problem)
- **Unclear relationships between foods and metabolites**



FOBI (Food-Biomarker Ontology)

Pol Castellano, Alex Sánchez

- Create an ontology that clearly defines the many complex relationships between **diet derived metabolites** and **foods** in a consistent and homogeneous way

- Reuse previous existing terms to maintain a consistent and standardized nomenclature (OBOFoundry)
  - FoodOn
  - ChEBI

- Propose a consistent starting point for nutrimetabolomic studies
  - Design
  - Validation

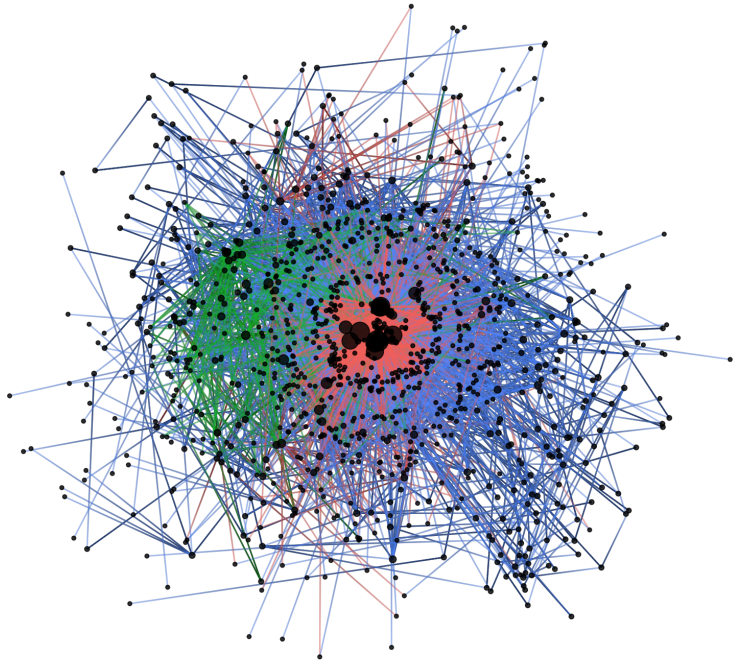
FOBI (Food-Biomarker Ontology)

Pol Castellano, Alex Sánchez

https://github.com/pcastellanoescuder/FoodBiomarkerOntology

Metrics

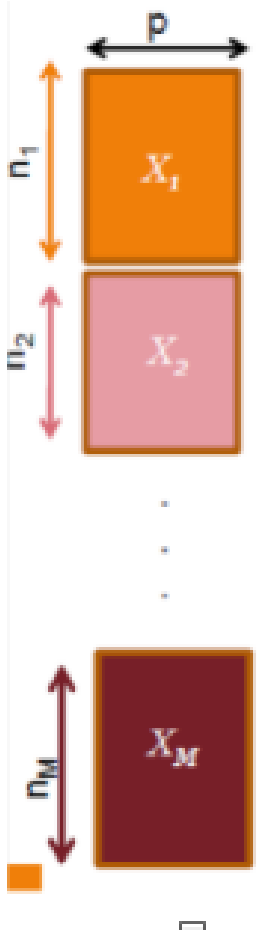
- 2 sub-ontologies
- 1197 terms
- 4 different properties
- 13 food top-level classes
- 11 biomarker top-level classes
- More than 4700 relationships
- Part of **OBOFoundry project** http://purl.obolibrary.org/obo/fobi.owl
- FOBI IDs are indexed into the HMDB (Human Metabolome Database) and FooDB (Food Database)



Insights in Multi-Group Data Analysis

Carolina Millapan, Ferran R., Esteban V.

- Analysis of variables observed in a set of individuals that belong to dierent groups.
  - Monitoring metabolite levels in blood of patients in different healthy/disease conditions,
- Measurements of heavy metals in soil samples from dierent ecological environments,
- Studies of gene expression for individuals with dierent experimental conditions



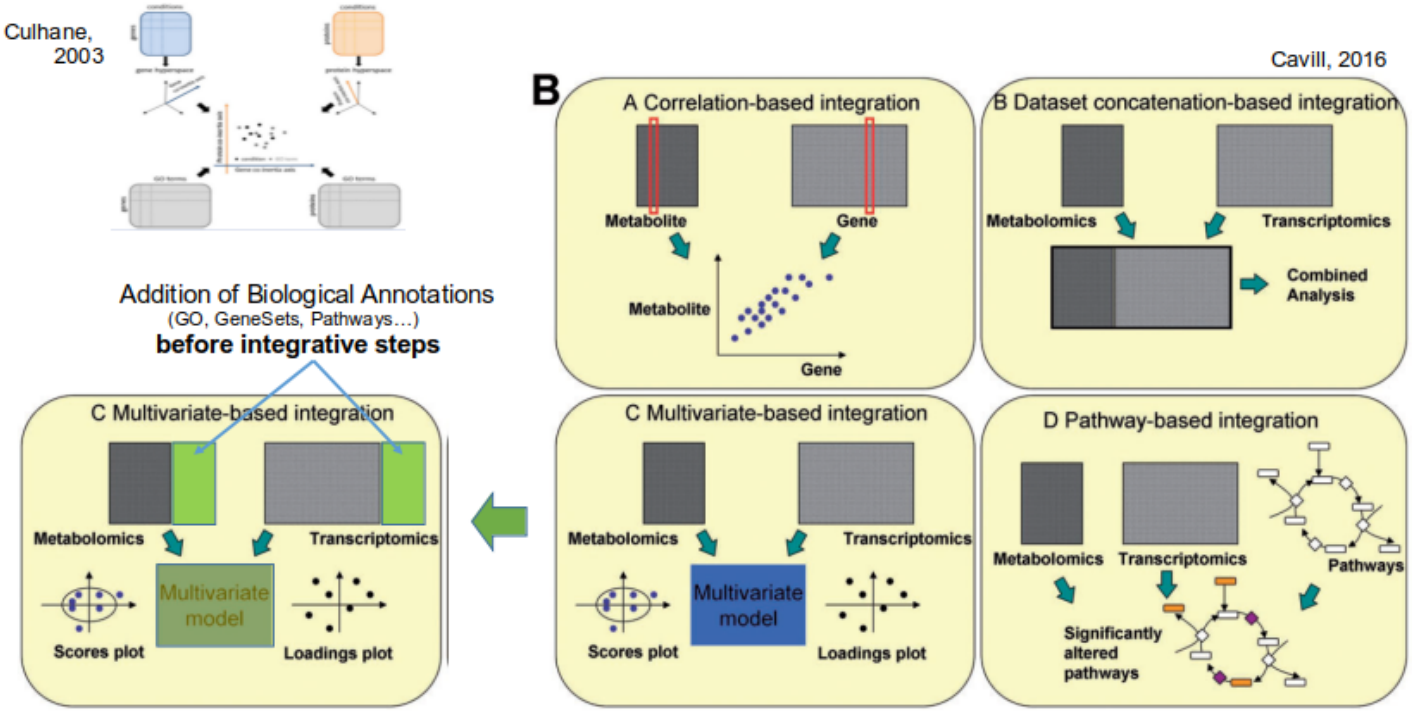
Carolina Millapan, Ferran R., Esteban V.

Multigroup PCA

- **Multigroup Principal Component Analysis**
  - Aim: *maximization of the interdistances between the points of the different groups*.
- **Multi-Group Dimension reduction**
  - Aim: find an orthogonal set of vectors such that each vector v tends to avoid that the projections of the observations of a given group do not overlap the centroids of the projections of the observations of each of the remaining groups.

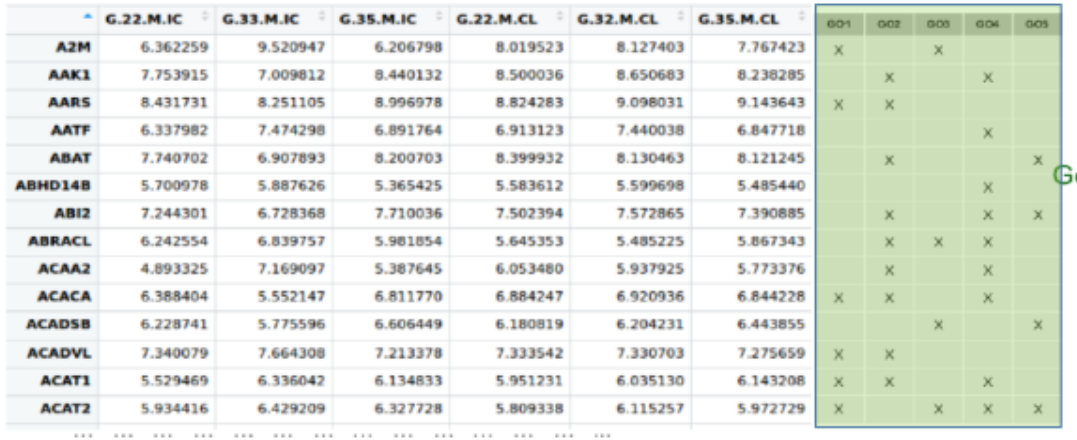
Data Integration with Annotations

Ferran Briansó, Alex Sánchez



Data Integration with Annotations

Ferran Briansó, Alex Sánchez



	G.22.M.IC	G.33.M.IC	G.35.M.IC	G.22.M.CL	G.32.M.CL	G.35.M.CL	GO1	GO2	GO3	GO4	GO5
A2M	6.362259	9.520947	6.206798	8.019523	8.127403	7.767423	x		x		
AAK1	7.753915	7.009812	8.440132	8.500036	8.650683	8.238285		x		x	
AARS	8.431731	8.251105	8.996978	8.824283	9.098031	9.143643	x	x			
AATF	6.337982	7.474298	6.891764	6.913123	7.440038	6.847718				x	
ABAT	7.740702	6.907893	8.200703	8.399932	8.130463	8.121245		x			x
ABHD14B	5.700978	5.887626	5.365425	5.583612	5.599698	5.485440				x	
ABI2	7.244301	6.728368	7.710036	7.502394	7.572865	7.390885		x		x	x
ABRACL	6.242554	6.839757	5.981854	5.645353	5.485225	5.867343		x	x	x	
ACAA2	4.893325	7.169097	5.387645	6.053480	5.937925	5.773376		x		x	
ACACA	6.388404	5.552147	6.811770	6.884247	6.920936	6.844228	x	x		x	
ACADSB	6.228741	5.775596	6.606449	6.180819	6.204231	6.443855			x		x
ACADVL	7.340079	7.664308	7.213378	7.333542	7.330703	7.275659	x	x			
ACAT1	5.529469	6.336042	6.134833	5.951231	6.035130	6.143208	x	x		x	
ACAT2	5.934416	6.429209	6.327728	5.809338	6.115257	5.972729	x		x	x	x

	G.22.M.IC	G.33.M.IC	G.35.M.IC	G.22.M.CL	G.32.M.CL	G.35.M.CL
SumGO1	39.986358	44.183758	...	...		
SumGO2	61.564510	...	...			
SumGO3	...	...				
SumGO4	...	...				
SumGO5	...	...				

$$\sum_{i_k \in A_k} x_{i_k j}$$

New variables from combination
of expression values and GO annotations

Different on each sample and each GO term

GENEONTOLOGY
Unifying Biology

Gene Enrichment Analysis
(hypergeometric tests
vs all GO categories)

(adjusted) P-values
< threshold

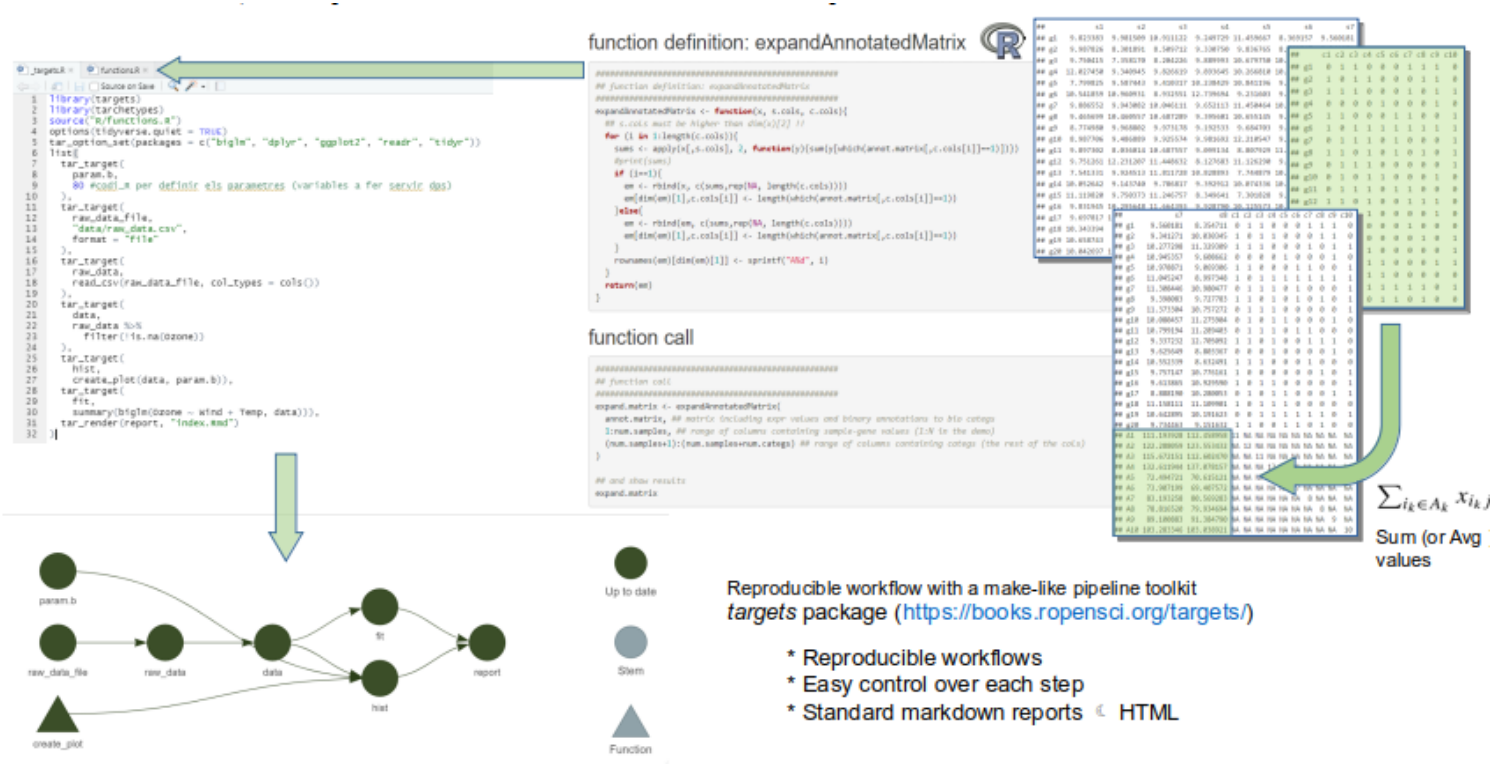
n Vars (genes)
m Samples (patients)
k Enriched Annotations (GO)

$$x_{ij} \quad \begin{matrix} i = 1..n \\ j = 1..m \end{matrix}$$

$$A_k \subset \{1..n\}$$

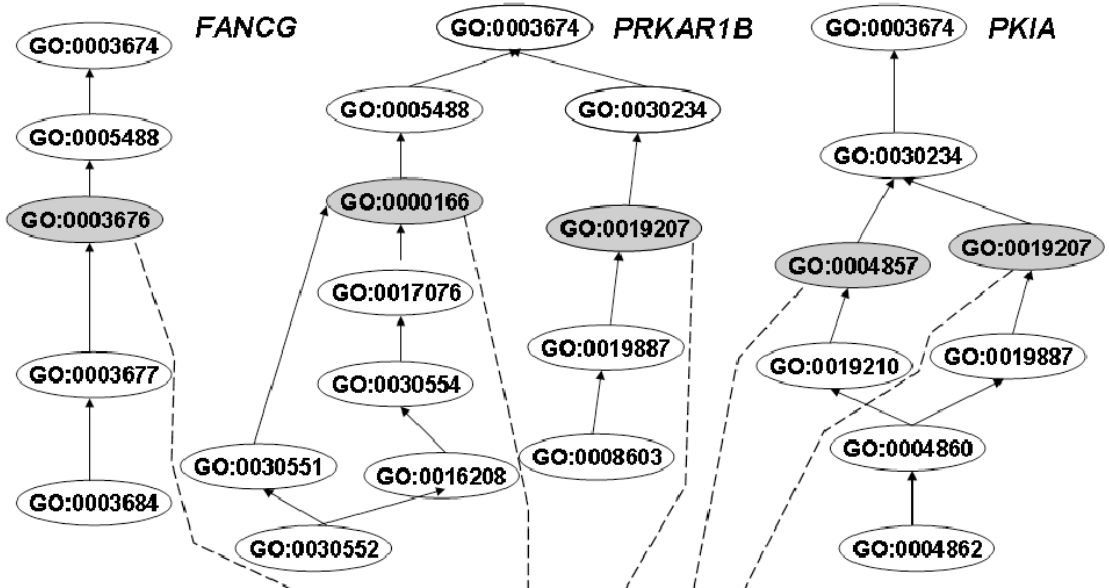
Data Integration with Annotations

Ferran Briansó, Alex Sánchez



The goProfiles Approach

Pablo Flores, Jordi Ocaña, Miquel Salicrú i Álex Sánchez



GO term\gene	FANCG	PRKAR1B	PKIA	*Profile*	%
GO:0003676	1	0	0	*1*	33,33%
GO:0000166	0	1	0	*1*	33,33%
GO:0019207	0	1	1	*2*	66,67%
GO:0004857	0	0	1	*1*	33,33%

Biomedical data integration and data sharing

Precision or Personalized Medicine

- Two patients with the same disease can respond very differently to the same treatment.
- Why? Mainly due to the different genomic characteristics of each one.
- Precision or Personalized Medicine aims at helping physicians to select the treatments that are most likely to help patients based on their genetics.
- PM is currently experiencing significant advances due to the appearance of new diagnostic and computer methods that provide an understanding of the molecular bases of the disease, particularly of the genomics.

Personalized Medicine at a Scale

The need for a paradigm change

- Until recently, PM has been associated with experimental therapies *far from daily clinical practice*
- But its central elements arae now much easier to reach:
    - Increasing availability of Genomic Data, fast and cheap to produce
    - Increasing computing power in-house or in the cloud (EOSC)
    - Possibility of accessing clinical data in anonymized possibly federated ways
- Having the ability to **integrate** genomic, clinical and other social, environmental and behavioral factors, will (should) lead to a transformation in the way in which decisions are made with direct effect on clinical practice and public health measures.
- The ultimate goal is a safer, more efficient, preventive, and predictive medicine.

How will this be done (in Spain)

The IMPaCT program

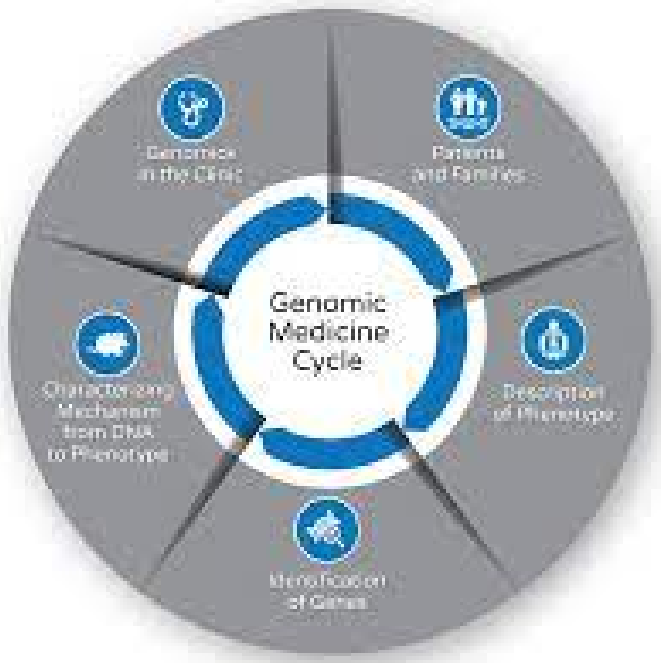
- In mid-September 2020, the Council of Ministers approved the call for granting grants for the **Precision Medicine Infrastructure associated with Science and Technology** (**IMPaCT**) of the Strategic Action in Health 2017-2020.
- This decision, which involved the granting of 25.8 million euros to the Carlos III Health Institute (ISCIII)
- The programs included in IMPaCT are aligned with several areas that will be developed in the future national strategy:
    - Predictive medicine;
    - Genomic medicine and
    - Data science.

The Predictive Medicine Program

- Aimed at the design and implementation of *a large population cohort with clinical, epidemiological and biological data, measured at the individual level*,
- which allows representing the entire population residing in Spain,
- including the ethnic variability and geographic and environmental diversity.
- **Data sharing** is an essential aspect: clinical data needs to be accessible for this program to be possible.

The Genomic Medicine Program

- It will develop *coordination infrastructures and protocols* to carry out genomic analyzes and other 'omic' data in an effective, efficient and equitable accessible manner.
- It will take as support large research centers that already have of state-of-the-art sequencing technology and experience in its application to the diagnosis of human diseases.
- It will optimize and reinforce the available massive sequencing capacities, orienting them to the needs of genetic diagnosis -exomes, complete genomes, etc.-

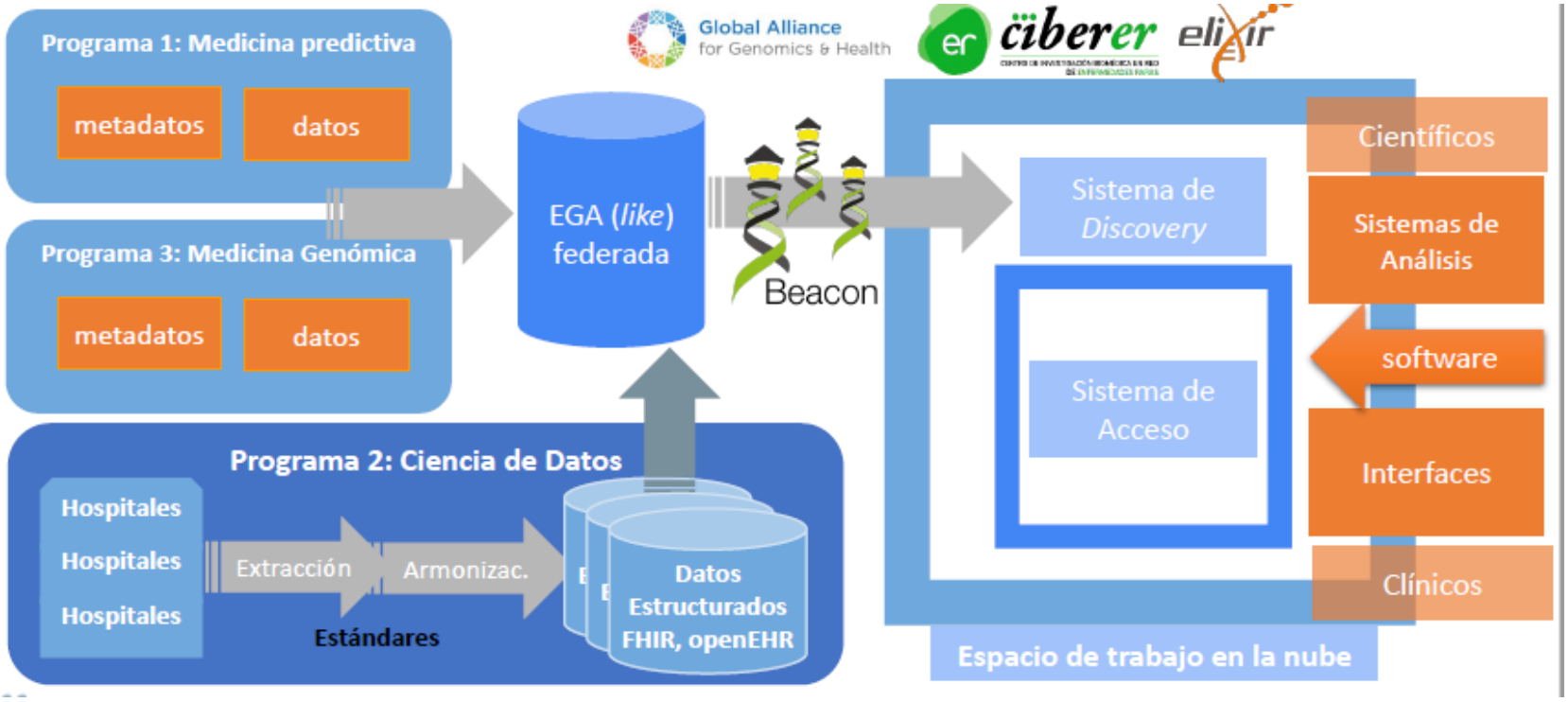


The Data Science Program

- *Develop a system for the collection, integration and analysis of clinical and molecular data_* aimed at improving the health of each individual patient, and
- Which allows the secondary use of existing information in the SNS to the benefit of society with the objectives of public health, health planning and research.
- Its objective is to **optimize the management of the information generated**, in order to apply it in the most effective way for the population and the SNS.
- It will provide
    - *bioinformatics tools* for the management of genomic data and
    - *medical informatics* solutions for the *management* and *integration* of clinical data,

facilitating the interoperability of the clinical information systems of the different autonomous health systems.

IMPaCT Global View



New Skills:

data sharing,

data exchange (OMOP, HL7)

data FAIRification,

cloud computing,

open science,

reproducibility,

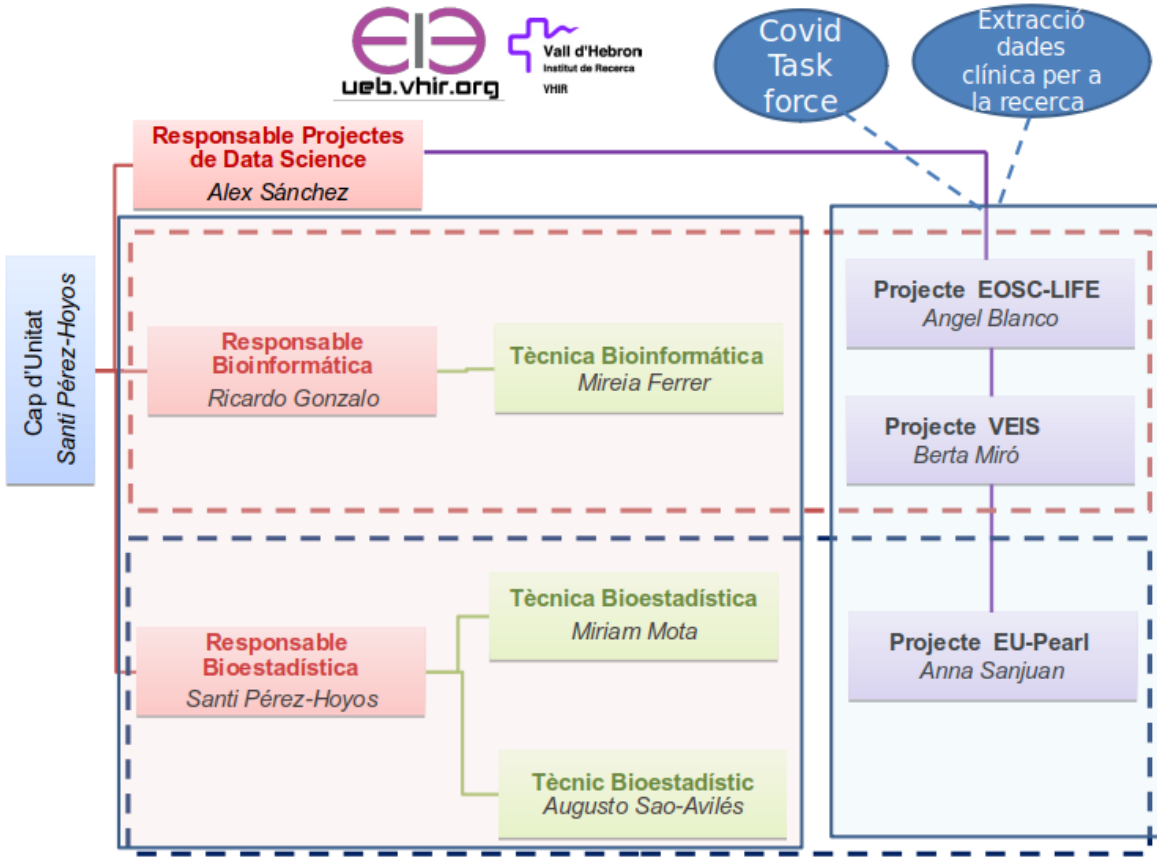
(federated) data analysis,

and, of course AI and ML

Data Projects we work in @VHIR

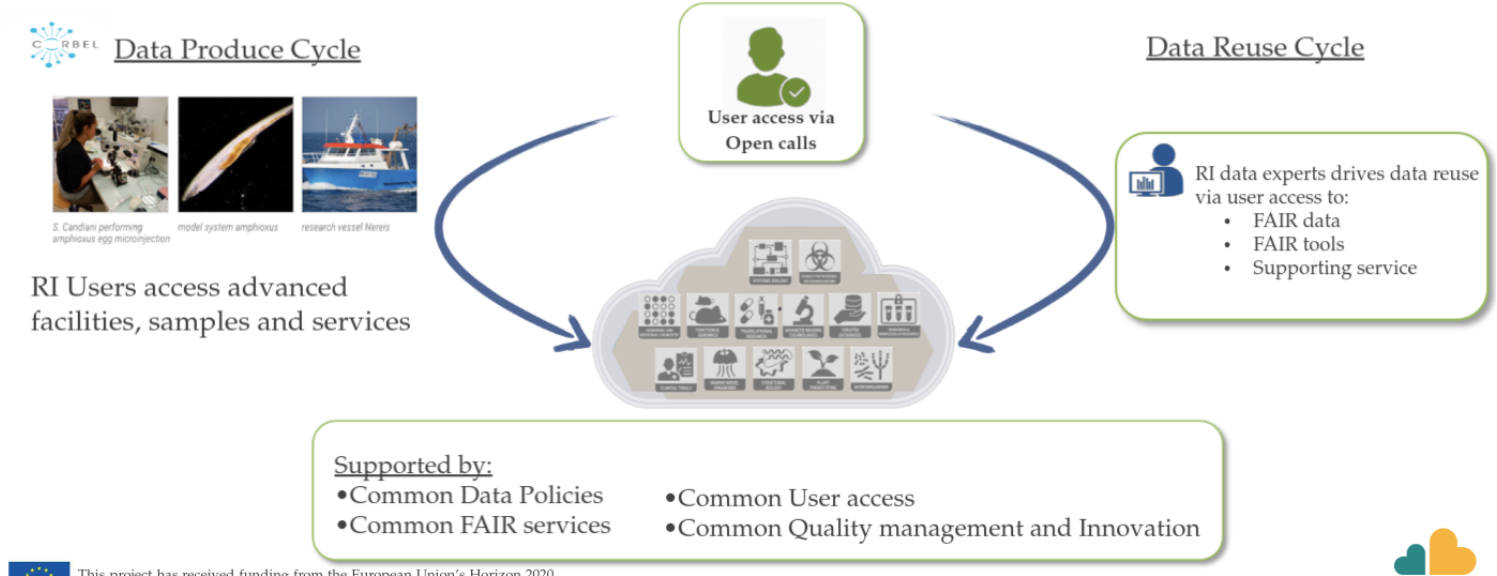
Data Science Projects Platform @VHIR

- The increasing needs in Data related projects leads to new structures to manage this.



External Data Projects: EOSC-Life

Enable ground-breaking data driven research in Europe by connecting life scientists to EOSC



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

EOSC-Life

External Data Projects: EU-PEARL

Vision

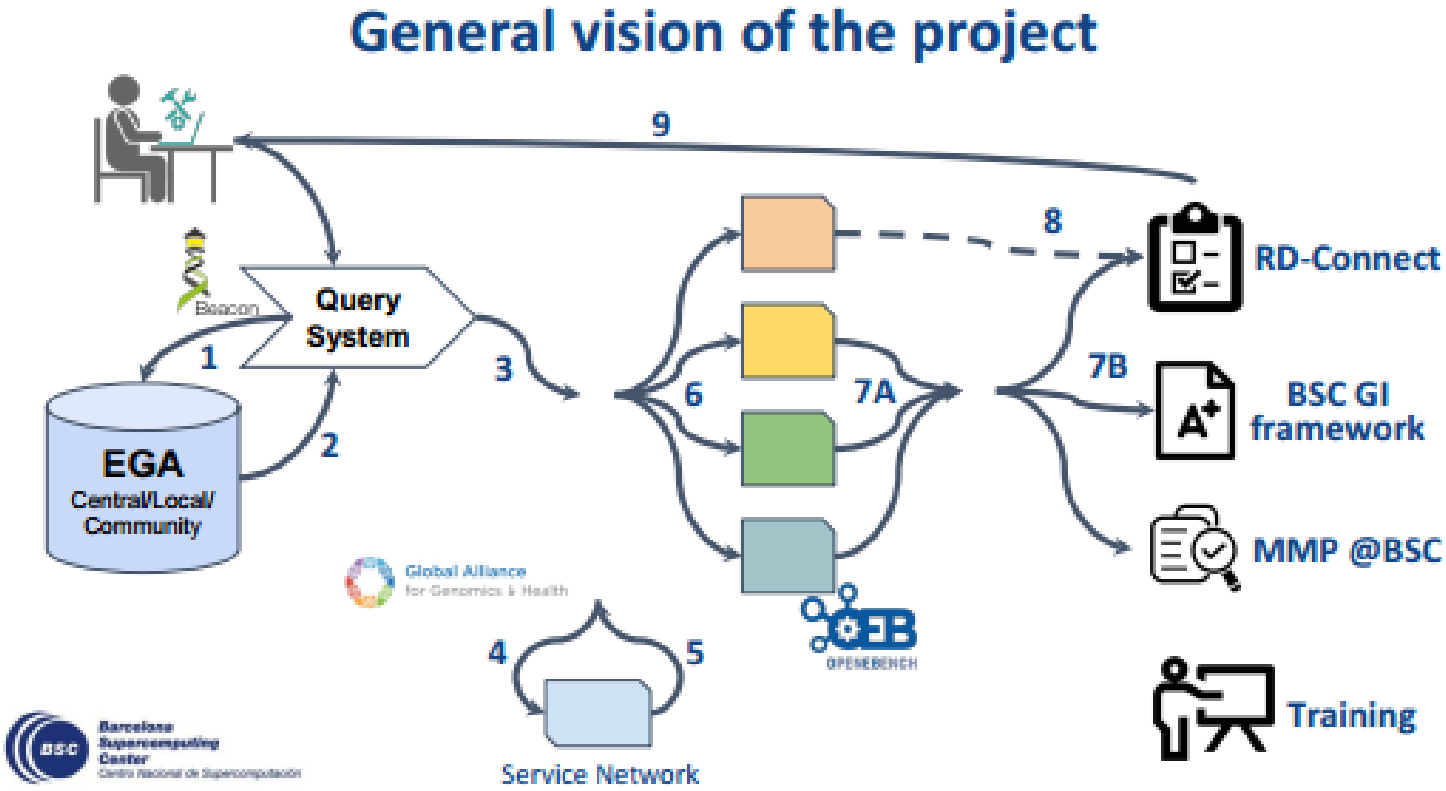
EU-PEARL aims to improve health outcomes for patients by shaping the clinical trials of the future. To achieve this vision, we have joined forces to create a sustainable and replicable framework that will produce a systematic approach to patient-centric, cross-company, multi-compound trial platforms.

Mission

EU-PEARL has set itself four objectives:

1. Create **a trusted, sustainable and replicable entity**, ready to setup and coordinate the operation of integrated research platforms (IRPs) in any disease area with unmet needs.
2. Set up an open, dynamic and patient inclusive IRP **Governance structure** to manage regulatory, ethical, legal, statistical and data requirements.
3. **Disseminate and exploit the EU-PEARL paradigm** by providing the necessary common tools, procedures, expertise and operational skills that meet the highest scientific, regulatory and ethical standards and best practices. These are to be developed jointly by public and industry partners in a consensus-based approach.
4. Create **trial ready IRP networks** to operate on **Major Depressive Disorder, Tuberculosis, Non-Alcoholic Steatohepatitis (NASH)** and **Neurofibromatosis**.

External Data Projects: VEIS



Other External Data Projects

- The European Health Data Space (TEHDAS)
- EHDEN (European HEalth Data and EVidence Network)
- DARWIN (Data Analysis and Real World Interrogation Network)

All these projects

- Are huge (all over Europe).
- Aim at some form of data re-use for health related research.
- Require different types od (trans-national) data sharing.
- With common or similar technical and legal issues to be addressed

Looking ahead

- The analysis of Omics Data is an example of fruitful development between Statistics and Biological (Health) Sciences.
- As more data is becoming available some therapeutical possibilities approach reality.
- While Statistics will always be relevant to guide this approach other skills become important.
- We must, at least, be aware of these skills and the problems they address because this is really an unprecedent opportunity.

Acknowledgements