

# Omics in human diseases

## Index

- [Omics data and Biological databases](#)
- NGS data analysis
- Prediction and interpretation of pathogenic variants
- Protein-protein interaction networks

### Course organization 2022/2023

**Monday:** Frontal lecture

**Thursday:** Frontal lecture/ guided practical activity

**How to pass the exam:** multiple choice quiz (50%) + **results** from practical activities (50%) + Bonus points, e.g. summary of previous lecture (up to 10%)

Mail: [emanuela.leonardi@unipd.it](mailto:emanuela.leonardi@unipd.it)

# Omics in human diseases

NASA video (1 of 8)

Introduction to Omics: 360 Degree View of You

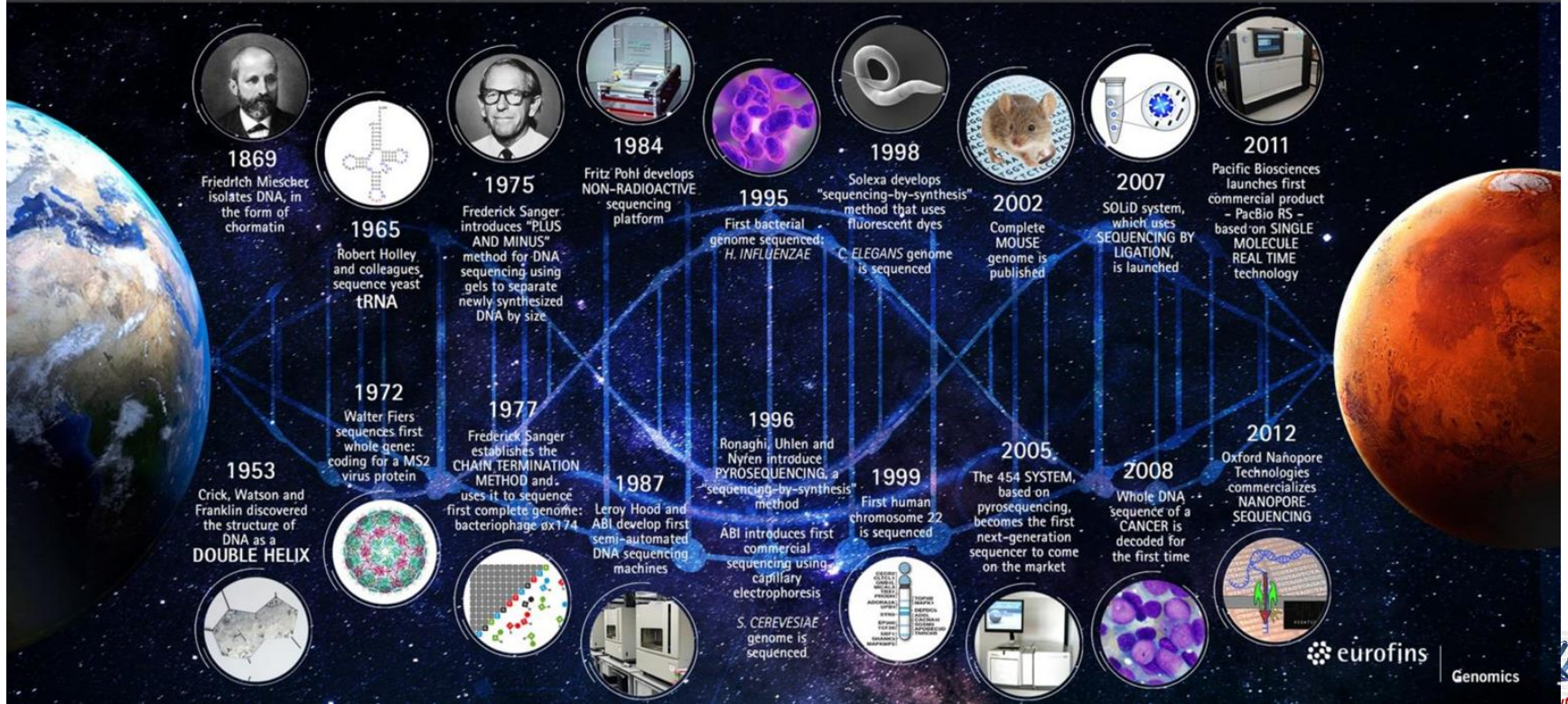
<https://www.youtube.com/watch?v=m7X6mugpijQ>

*Listen and answer*

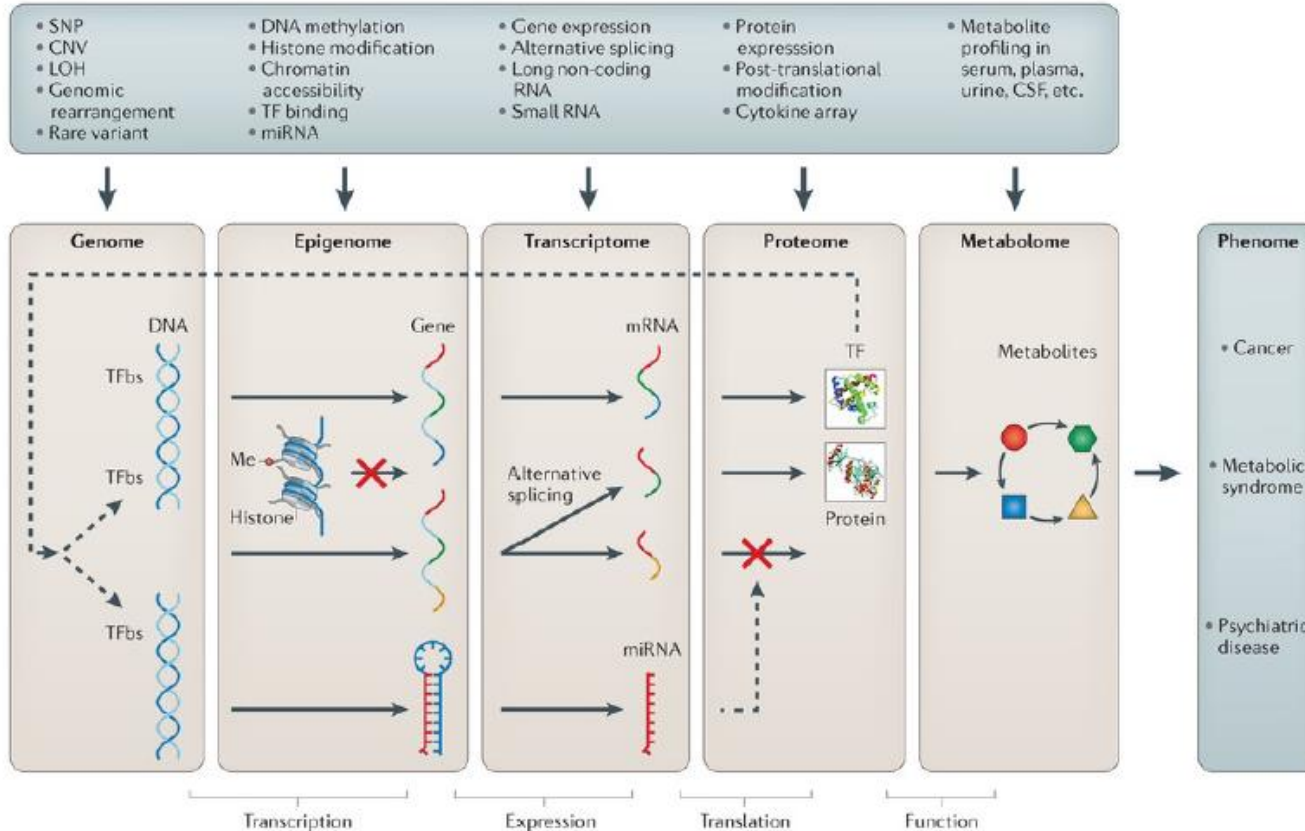
1. How “Omics” can improve our health system?
2. Which metaphor can we use to describe the role of “Omics”?

# Technological revolution

## A JOURNEY THROUGH THE HISTORY OF DNA SEQUENCING

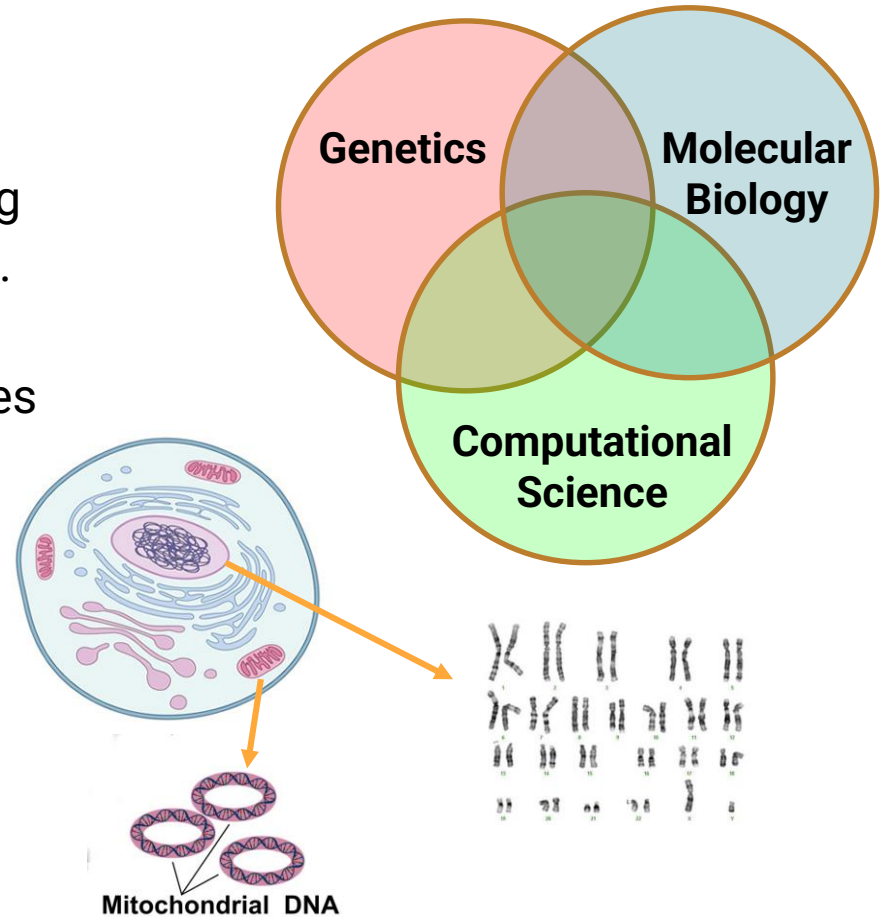


# Omics in human diseases



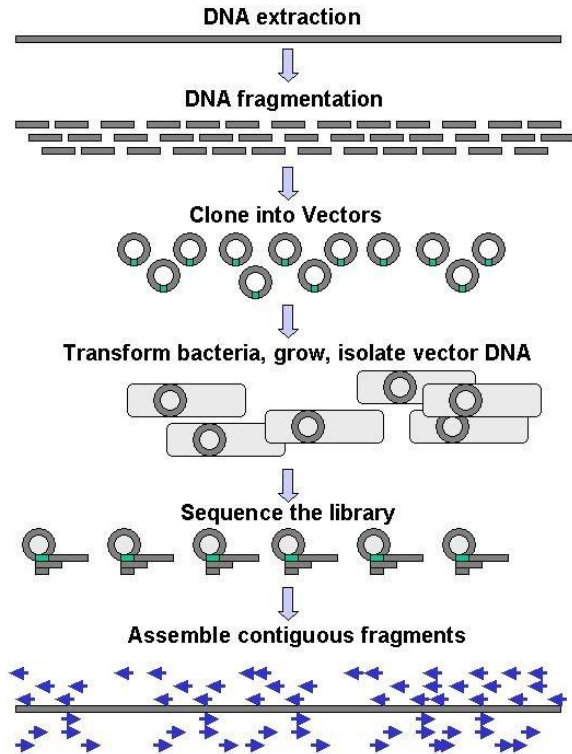
# Genomics

- The study of complete Genomes including their organization, **function**, and evolution.
- **Human genome**: 23 pairs of chromosomes in the nucleus, and mitochondrial DNA (coding and non-coding sequence)
- Mapping and studying **genetic variants** associated with diseases, response to treatment, or future patient prognosis.

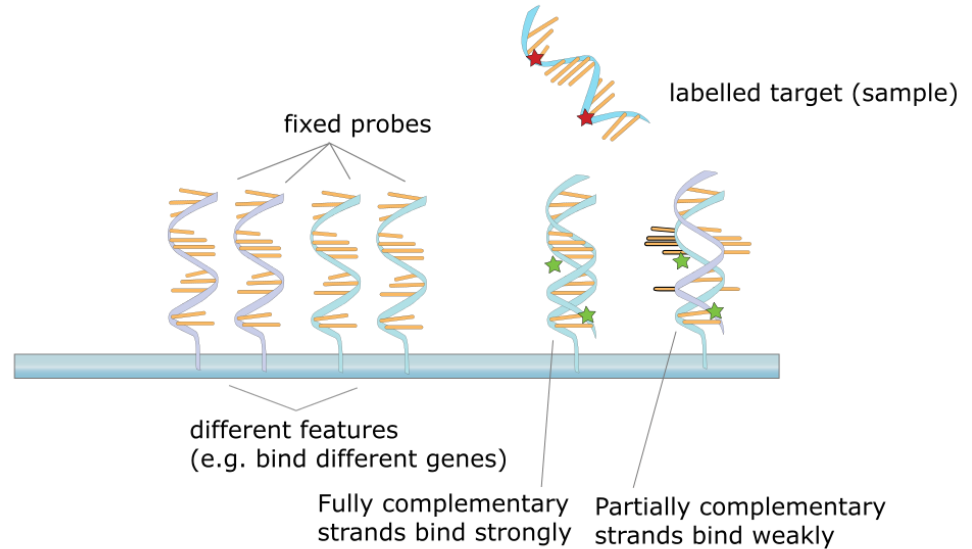
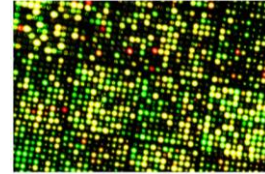


# Genomics: Technology

## Massive parallel sequencing



## DNA microarray/SNP genotyping



# Catalog of Genome Wide Association Studies (GWAS)

<https://www.ebi.ac.uk/gwas/>



**GWAS Catalog**

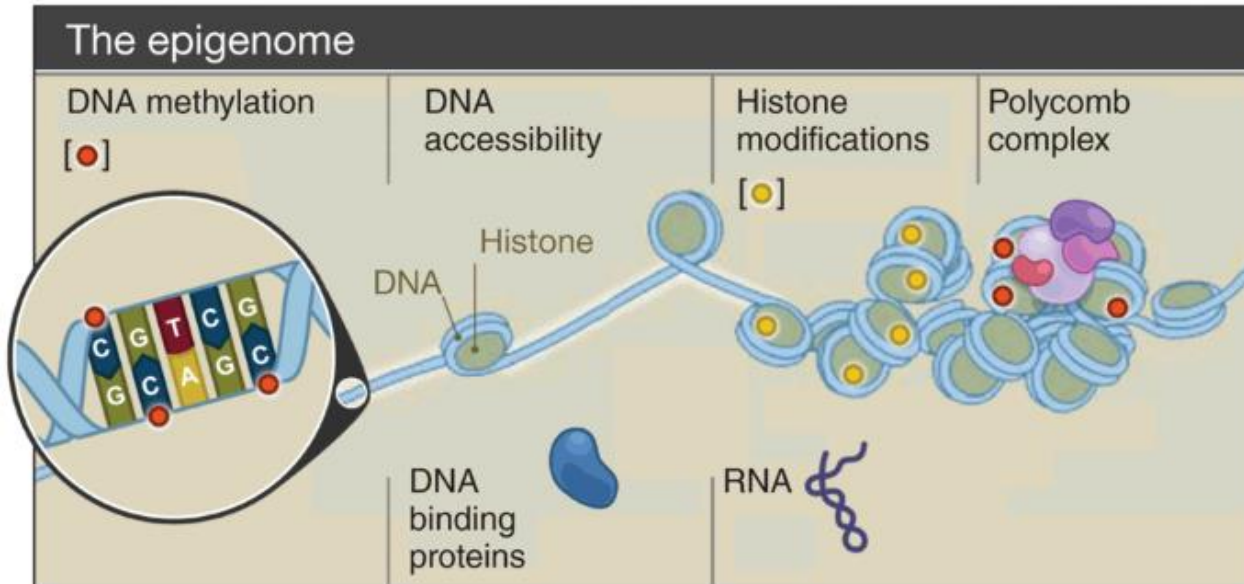
The NHGRI-EBI Catalog of human genome-wide association studies

Q

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

- Database of SNP-trait associations
- Integrated with other resources
- Accessible by scientists, clinicians and others

# Epigenomics



The term 'epigenome' focuses on genome-wide characterization of reversible modifications of DNA and DNA-associated proteins

- Influenced by genetics and environment
- Tissue- specific
- Differentially methylated regions of DNA (Epigenetic signatures) as indicators of disease
- Functional interpretation of genetic variants in differentially methylated regions



# Epigenomics: <http://www.roadmapepigenomics.org/>



Search:  GO

HOME

PARTICIPANTS

BROWSE DATA

PROTOCOLS

COMPLETE EPIGENOMES

TOOLS

PUBLICATIONS



OVERVIEW



PROJECT DATA



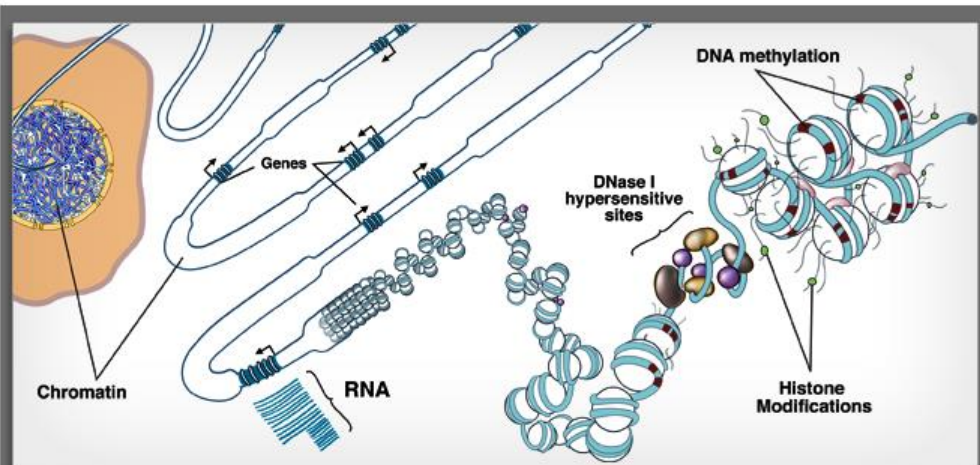
MAPPING CENTERS



PROTOCOLS & STANDARDS

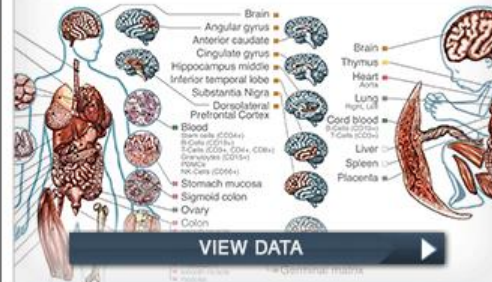


PUBLICATIONS



NIH Roadmap Epigenomics Mapping Consortium

## INTEGRATIVE ANALYSIS of 111 REFERENCE HUMAN EPIGENOMES



VIEW DATA

VIEW/DOWNLOAD QUICK LINKS

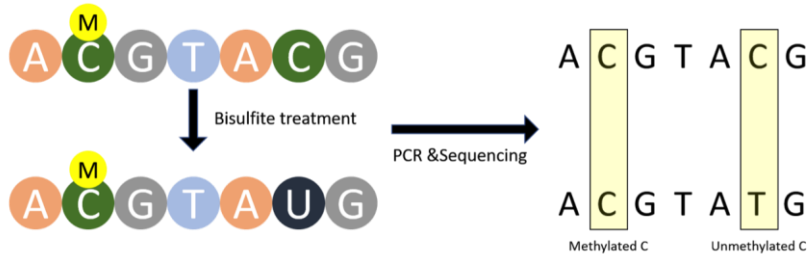
Genome Browsers

• <http://genomebrowser.wustl.edu/>

# Epigenomics: Technology

Assessment of DNA modification using NGS

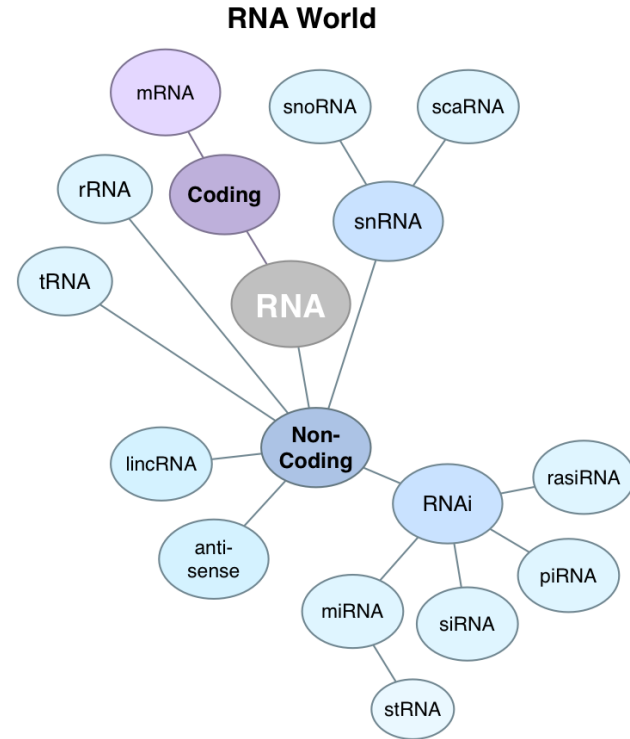
## Whole-genome shotgun bisulfite sequencing



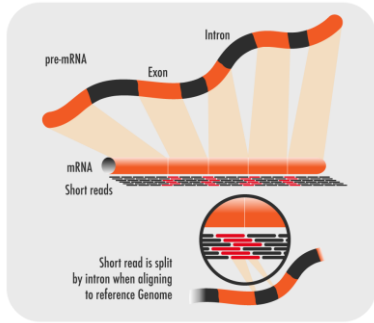


# Transcriptomics

- Genome-wide analysis of all RNAs under different cells, tissue, developmental stage, experimental or pathological conditions.
- Transcriptome include different types of RNA molecules
- Quantitative (transcripts level)
- Qualitative (which transcript, novel isoforms)



# Transcriptomics: technology

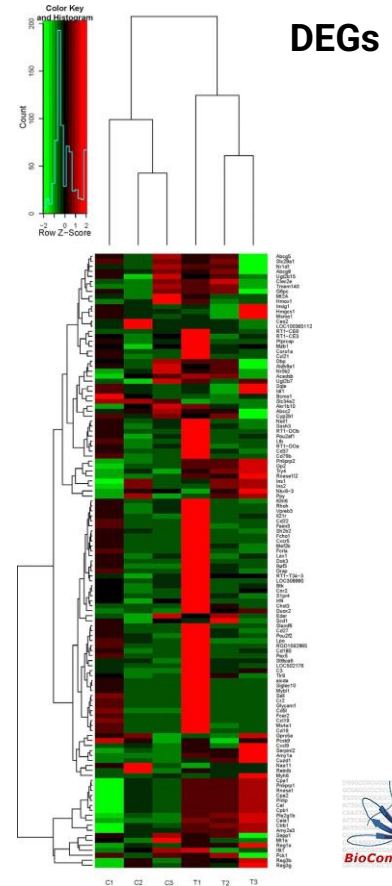


- RNA sequencing

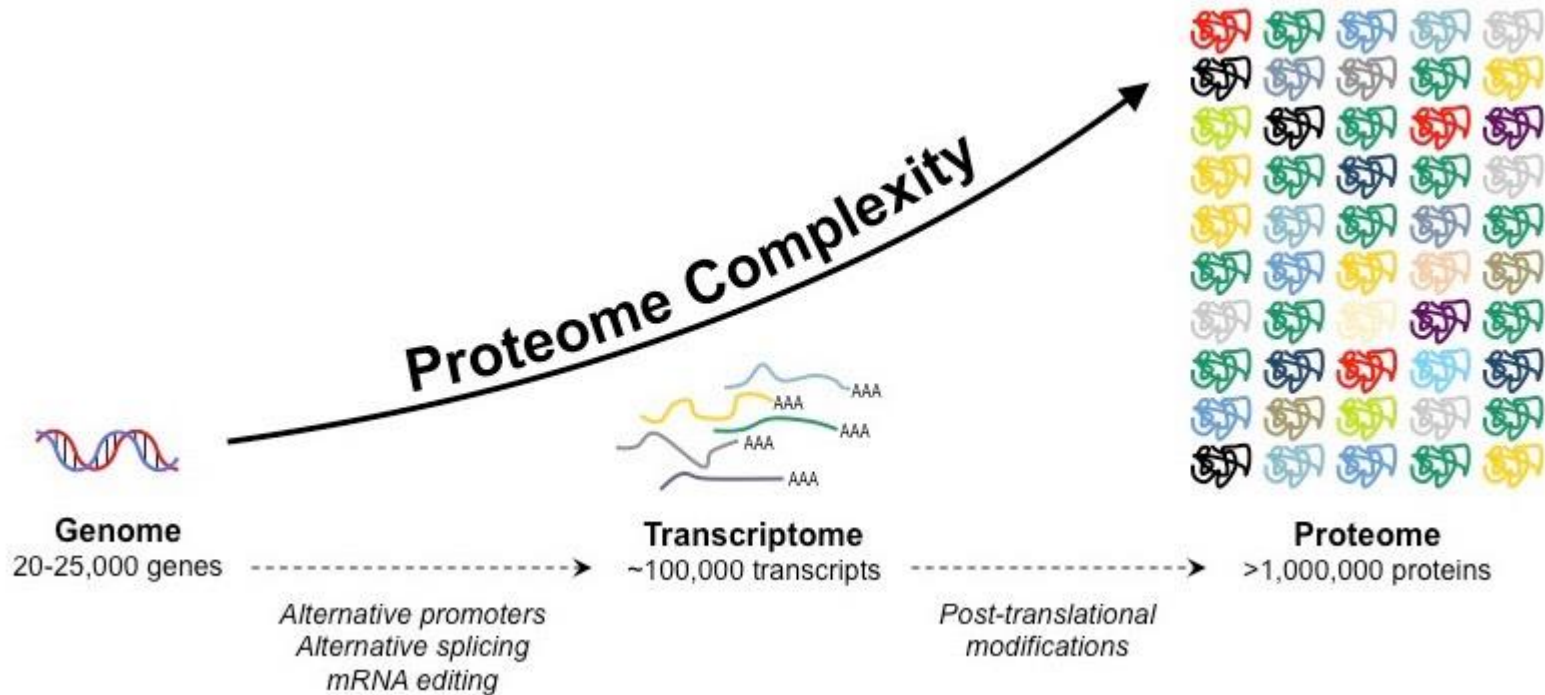
or

- Microarrays

RNA isolation and  
conversion into  
complementary  
DNA



# Proteome



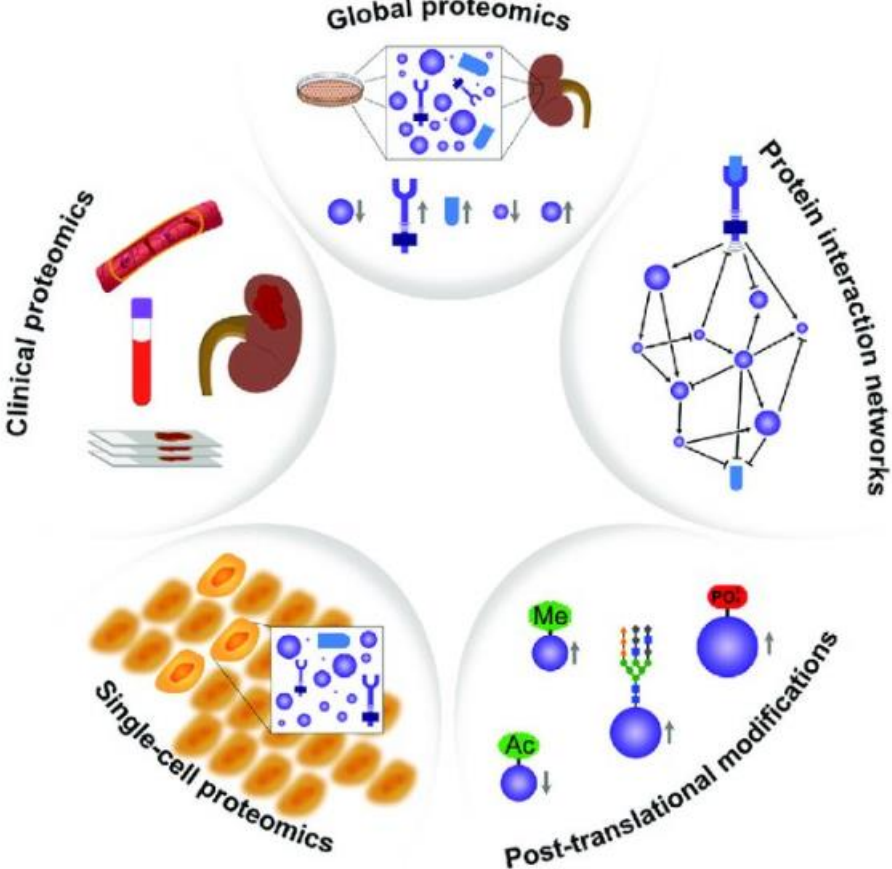
# Proteomics

- Large-scale systematic study of proteomes
- Proteome: The complete set of proteins expressed by an organelle, cell, tissue or organism at a certain time.

## Limitations

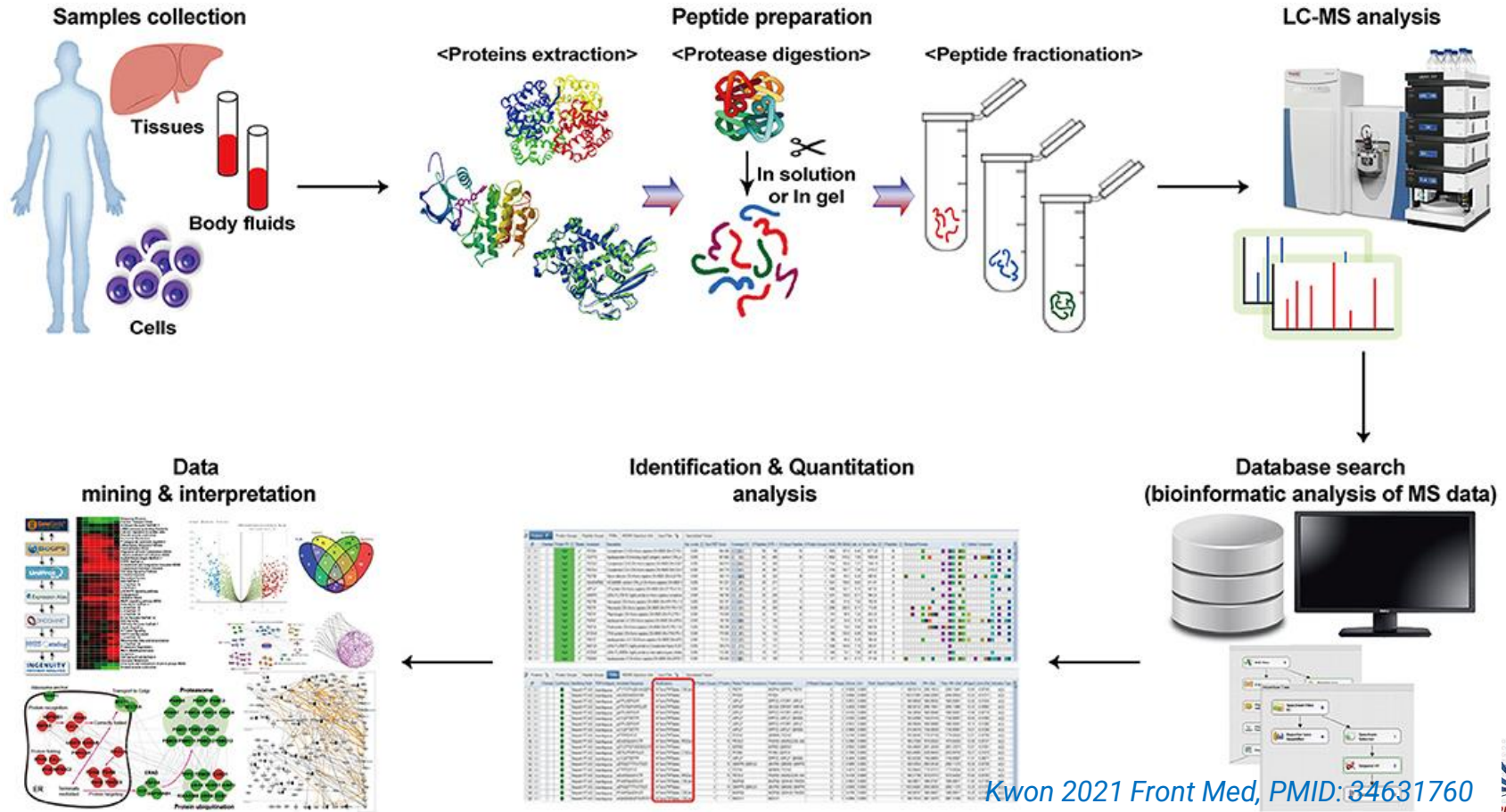
- PTM affect protein activity
- Alternative splicing or PTM give rise to more than one protein
- Many proteins form complexes with other proteins or RNA molecules
- protein degradation rate plays an important role in protein content

# Proteomics: applications



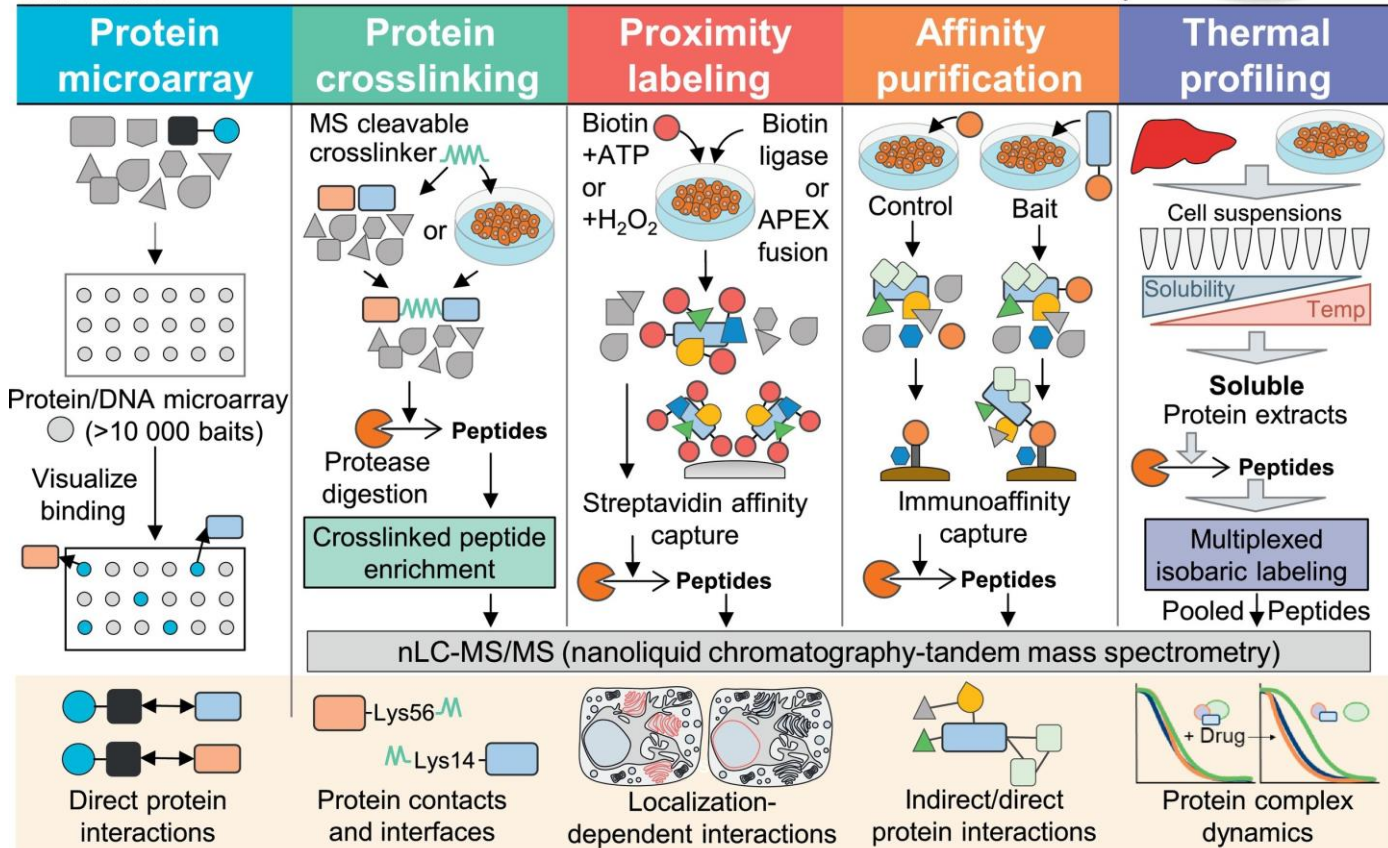
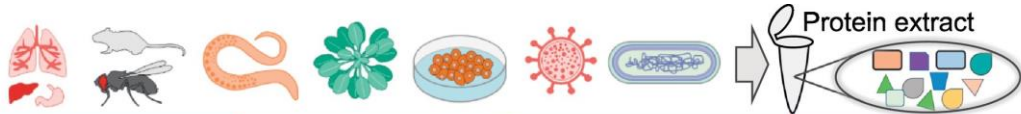


# Proteomics: workflow



# Proteomic Technologies for Deciphering Local and Global Protein Interactions

Animals, plants, viruses, bacteria cell culture



PMID: 32035732

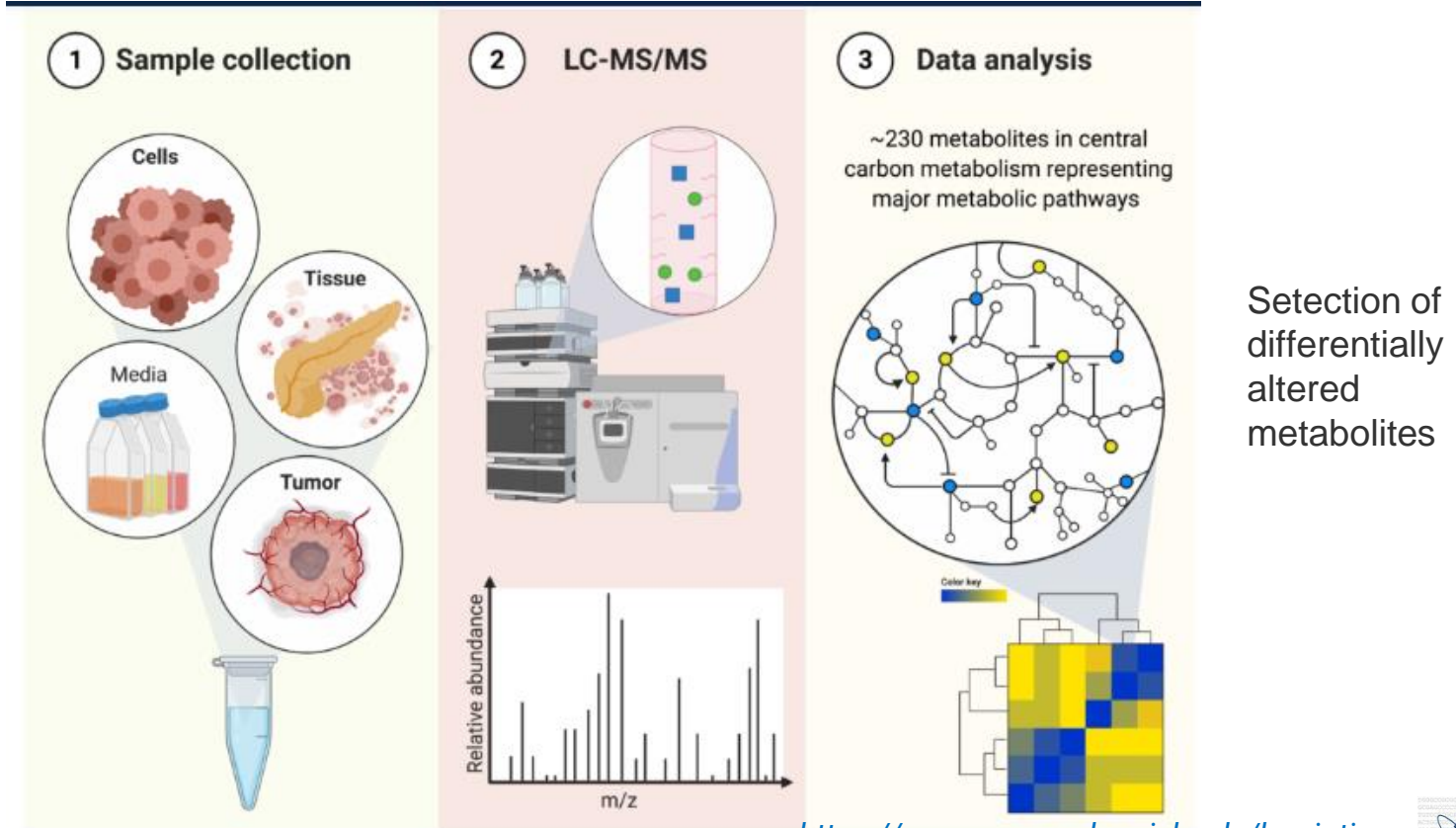
# Metabolomics

The **metabolome** represents the complete set of metabolites in a biological cell, tissue, organ, or organism, which are the end products of cellular processes

## Applications

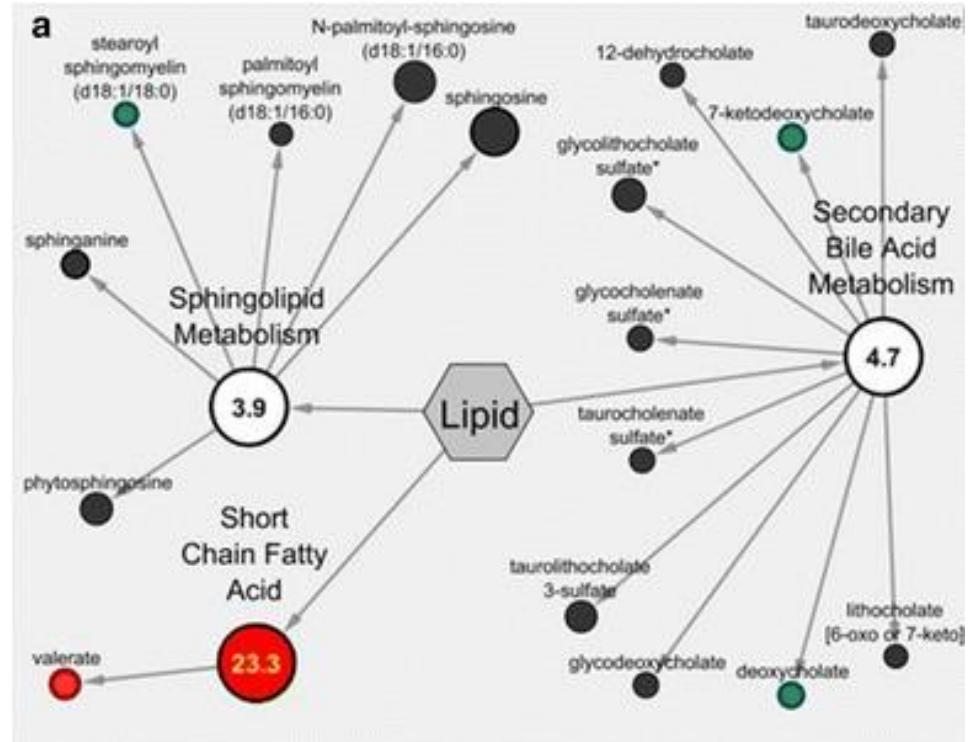
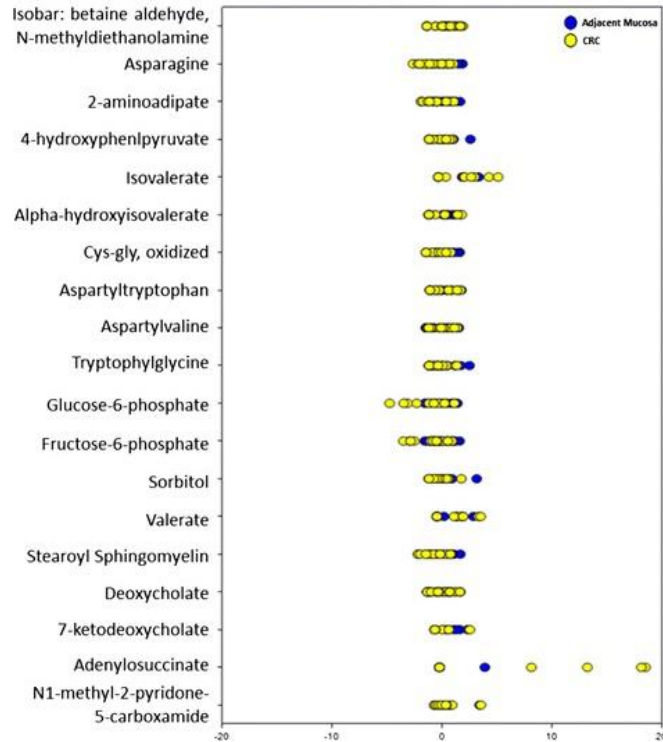
- Discover new, metabolomics-based biomarkers with high chances of translation into precision medicine
- Functional genomics: Identify the phenotype caused by a genetic alteration

# Metabolomics: workflow



<https://pancreas.med.umich.edu/lyssiotis-lab/research/metabolomics-research/>

# Metabolomics



Red nodes represent metabolites and pathways with higher expression in CRC

# Newborn metabolic screening

<https://www.iss.it/web/iss-en/newborn-screening>



- **Preventive** public health program
- **Blood** sample taken from each newborn baby's heel
- Test **free of charge**, 48 and 72 hours after birth, at the hospital
- In 2016, Italian Law No. 167 extended the newborn screening programme to include around **40 genetic disorders**
- For each of these diseases, a **therapeutic treatment** capable of improving longevity and quality of life now exists and is available.

# Phenomics

The study of sets of traits belonging to an organism.

The acquisition of high-dimensional phenotypic data on an organism-wide scale.

## Aim

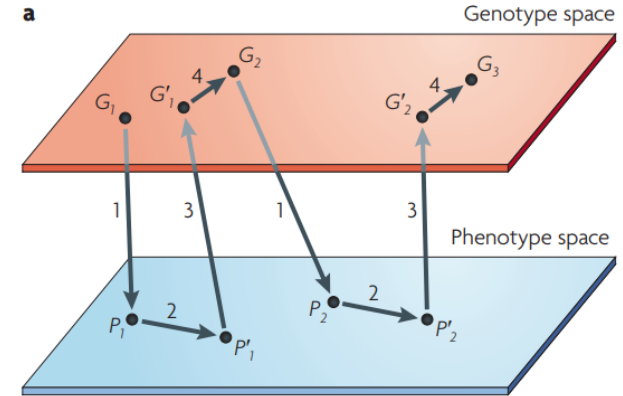
Better understanding of the **genotype–phenotype map**

genotypes (**G space**) and phenotypes (**P space**)

1. epigenetic process
2. natural selection act to change phenotype parents
3. Preserved phenotype, the identity of successful parents
4. Genetic events

## Challenges

- Phenotypes vary from cell to cell and from moment to moment and therefore can never be completely characterized.
- Development and adoption of high-throughput and high-dimensional phenotyping



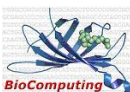
*Houle D, 2010 Nat Rev Genet PMID: 21085204*

# Phenomics

Phenomic projects that combine genomic data with data on quantitative variation in phenotypes

Description	Funding	Phenotypes	Genotyping
Consortium for Neuropsychiatric Phenomics. 52 investigator, interdisciplinary effort. Genomic data, brain structure and function and behaviour in case-control study of three major psychiatric syndromes ( <a href="http://www.phenomics.ucla.edu">http://www.phenomics.ucla.edu</a> )	NIH	Brain imaging, behaviour and cognitive phenotypes	Northern Finland Birth Cohorts, case-control genotyping
UK Biobank. Prospective study of 500,000 individuals ( <a href="http://www.ukbiobank.ac.uk">http://www.ukbiobank.ac.uk</a> )	MRC, Department of Health, Wellcome Trust	Baseline questionnaire and physical measurements; storage of blood and urine for eventual analysis and integration with the UK NHS health records	Samples taken for later analysis
Personal Genome Project: recruit volunteers for genome sequencing and phenotype data. Participant number: 100,000 as a goal ( <a href="http://www.personalgenomes.org">http://www.personalgenomes.org</a> )	Private	Images, cell lines and medical history	Primary goal is genome sequencing. One participant fully sequenced

*Houle D, 2010 Nat Rev Genet PMID: 21085204*



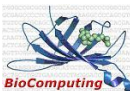


# Phenomics: Human Phenotype Ontology (HPO)

The HPO, as a part of the Monarch Initiative, is a central component of one of the [13 driver projects](#) in the [Global Alliance for Genomics and Health \(GA4GH\) strategic roadmap](#).

- Standardized vocabulary of phenotypic abnormalities encountered in human disease
- HPO terms can be used to describe the phenotypic features that occur in individuals with a disease
- Sources: medical literature, Orphanet, DECIPHER, and OMIM
- 13,000 terms and over 156,000 annotations to hereditary diseases. Initially (2005) focused on Mendelian diseases, extended on common disease in 2015.

*Kohler, 2021, NAR, PMID: 33264411*



# Phenomics: HPO (<https://hpo.jax.org/app/>)

Tools ▾ Downloads ▾ Documentation ▾ Di... ▾ autism

No. Descendants Hierarchy ?

- Behavioral abnormality
  - Autistic behavior
    - Impaired social interactions
    - Autism with high cognitive abilities
    - Autism
    - Restrictive behavior
    - Alexithymia

## Autistic behavior HP:0000729

*Persistent deficits in social interaction and communication and interaction as well as a markedly restricted repertoire of activity and interest as well as repetitive patterns of behavior.*

**Synonyms:** *Autistic behaviors, ASD, Pervasive developmental disorder, Autistic behaviour, Autism spectrum disorders, Autistic behaviours, Autism spectrum disorder*

**Comment:** This term can be used to refer to autism spectrum disorder as a phenotypic feature that can be a component of a disease. Autism spectrum disorder range from a severe form, called autistic disorder, to a milder form, Asperger syndrome.

**Pubmed References:** [PMID:28879490](#)

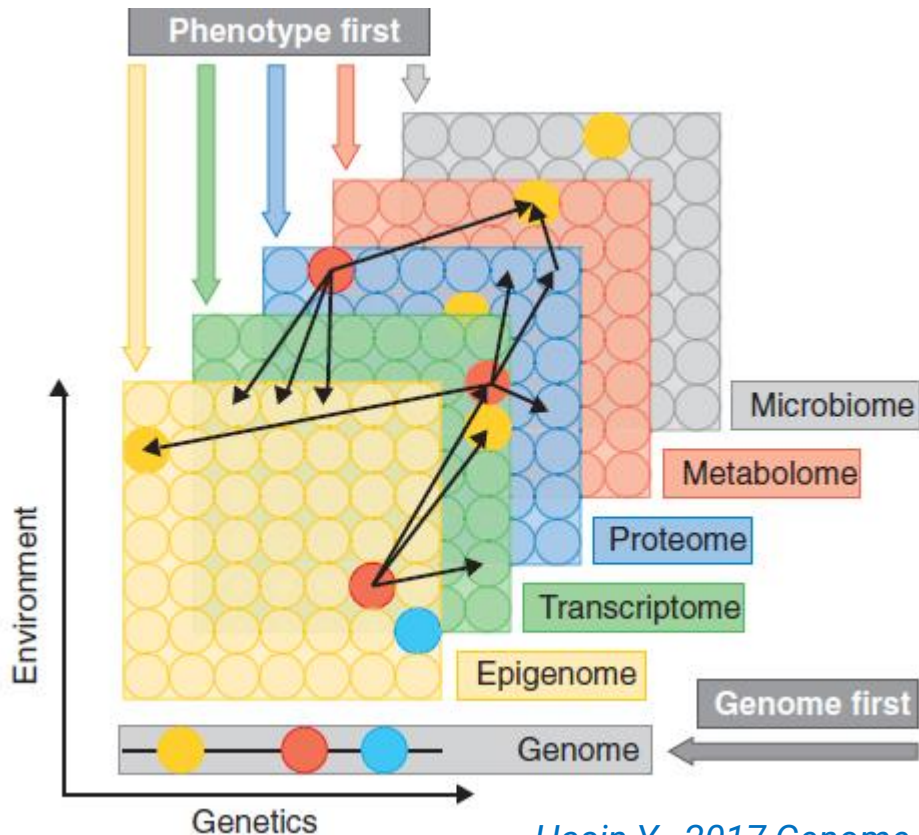
**Cross References:** [UMLS:C0856975](#), [UMLS:C1510586](#), [MSH:D000067877](#)

Export Associations

**Disease Associations** Gene Associations

Disease Id	Disease Name	Associated Genes
------------	--------------	------------------

# Multi-Omics approach



- Except for the genome, all **data layers** reflect both **genetic regulation and environment**, which may affect each individual molecule to a different extent.
- **The thin red arrows** represent potential interactions or correlations detected between molecules in different layers
- **Thicker arrows** indicate different potential starting points or conceptual frameworks for consolidating multiple omics data to understand disease.
- **Genome first approach** implies that one starts from associated locus, while the **Phenotype first approach** implies any other layer as the starting point.

Hasin Y., 2017 *Genome Biol.* PMID: 28476144

# Biological Databases

© 1997 by Randy Glasbergen.  
E-mail: randyg@norwich.net



**“I hope you’ve got a lot of disk space, Ted.  
I think I accidentally just faxed you  
the entire Internet.”**

# WHAT is a database?

A collection of data that needs to be:

- Structured
- Searchable
- Updated (periodically)
- Cross referenced

**Challenge:** To change “meaningless” data into useful information that can be accessed and analysed the best way possible.



wiseGEEK

# Databases are organized collections of information

Databases assign each record a unique **accession number** using their own numbering system



**Accession #**

**Fields** are used to cross-reference the data.

Records can be searched by fields.



**Field 1 .....**

**Field 2 .....**

**Data** is entered in the record using a defined format



**Data**

.....

.....

.....

**Bioinformaticians** work with computer scientists to set up the database structure

**Curators** review and link records within and between databases

# The information in databases ultimately derives from experimental data



Researchers do experiments



Researchers analyze data and write papers



Data are published in journals



PubMed

Curators will process the submissions and link entries in different databases



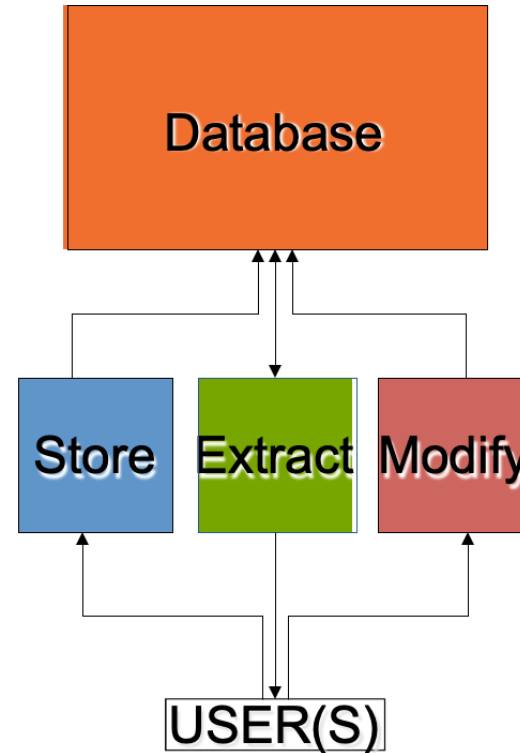
# Database management system (DBMS)

## Internal organization

→ Controls speed and flexibility

A unity of programs that

- Store
- Extract
- Modify





# DBMS organisation types

## Flat file databases (flat DBMS)

→ Simple, restrictive, table

## Hierarchical databases (hierarchical DBMS)

→ Simple, restrictive, tables

## Relational databases (RDBMS)

→ Complex, versatile, tables

## Object-oriented databases (ODBMS)

→ Complex, versatile, objects

## Data Warehouses and Distributed Databases



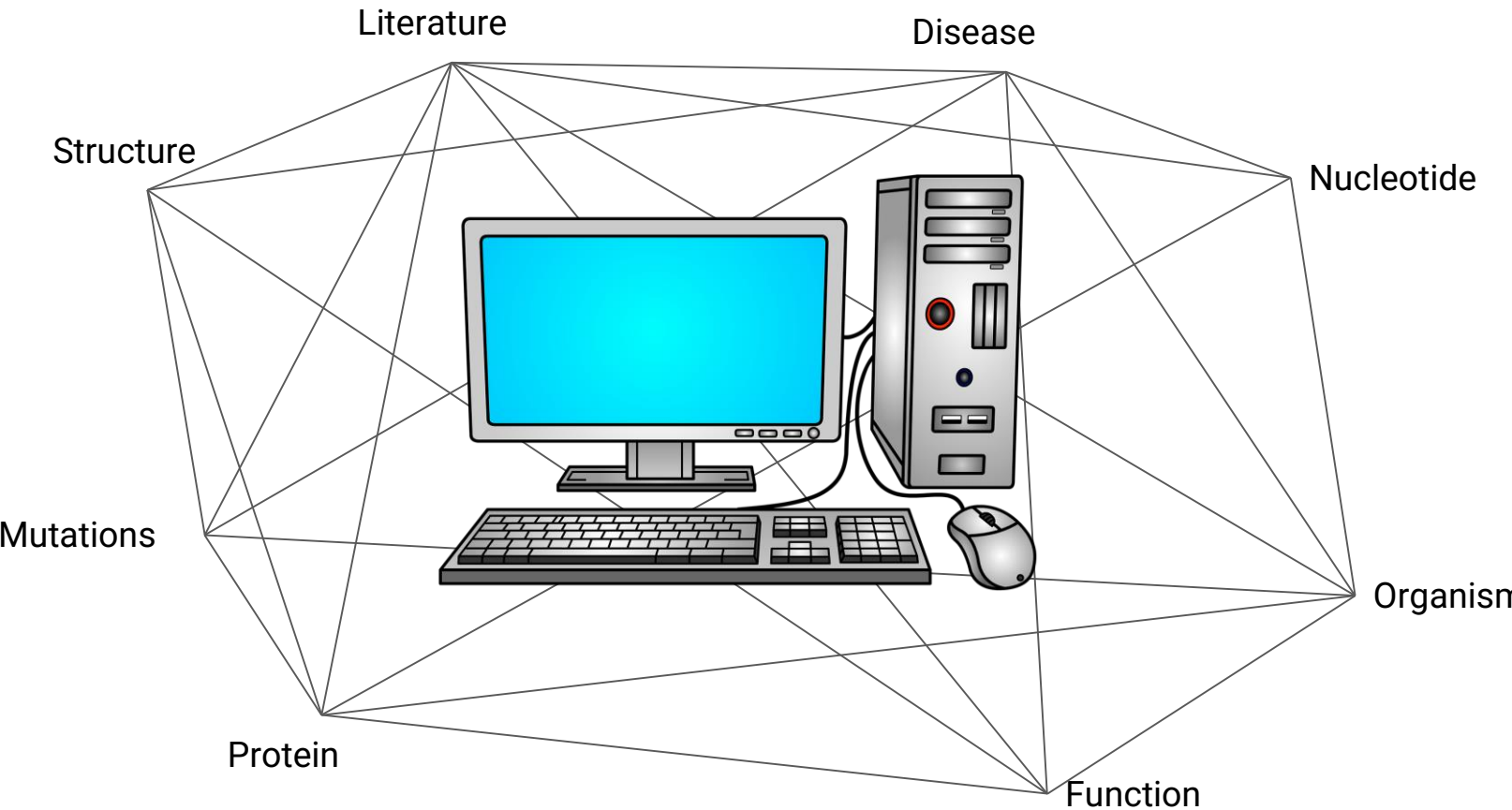
Information system

Query System

Storage System

Data

# Biological Databases

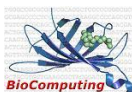


# Biological databases: why? & which types?

- Need for storing and communicating large datasets has grown
- Make biological data available to scientists
- To make biological data available in computer-readable form

## Type of data

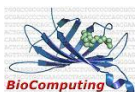
- nucleotide sequences
- protein sequences
- Genetic variants
- gene expression data
- Protein interactions
- metabolic pathways



# Biological databases

## Availability

- Publicly available, no restrictions
- Available, but with copyright
- Accessible, but not downloadable
- Academic, but not freely available
- Proprietary, commercial; possibly free for academics



# Standard Data Formats

DNA sequence = **ACGT**, but what about gaps, unknown letters, etc.

- How many letters per line ???
- Spaces, numbers, headers, etc. ???
- Store as a string, code as binary numbers, etc.

Use a completely different format for proteins?

# Standard Data Formats

DNA sequence = **ACGT**, but what about gaps, unknown letters, etc.

- How many letters per line ???
- Spaces, numbers, headers, etc. ???
- Store as a string, code as binary numbers, etc.

Use a completely different format for proteins?

Need standard formats!!

# FASTA Format

- William Pearson (1985)
- The FASTA format is now **universal** for all databases and software that handles DNA and protein sequences

One header line, starts with > with a [return] at end

→ All other characters are part of sequence.

```
>UR01 urol.seq Length: 2018 November 9, 2000 11:50 Type: N Check: 3854 ..
CGCAGAAAGAGGAGGCGCTTGCCTTCAGCTTGTGGGAAATCCCGAAGATGGCCAAAGACA
ACTCAACTGTTTCGTTGCTTCCAGGGCCTGCTGATTTTTGGAAATGTGATTATTGGTTGTT
GCGGCATTGCCCTGACTGCGGAGTGCATCTTCTTTGTATCTGACCAACACAGCCTCTACC
CACTGCTTGAAGCCACCGACAACGATGACATCTATGGGGCTGCCTGGATCGGCATATTTG
TGGGCATCTGCCTCTTCTGCCTGTCTGTTCTAGGCATTGTAGGCATCATGAAGTCCAGCA
GGAAAATTCTTCTGGCGTATTTCAATCTGATGTTTATAGTATATGCCTTTGAAGTGGCAT
CTTGTATCACAGCAGCAACACAACAAGACTTTTTACACCCAACCTCTTCTGAAGCAGA
TGCTAGAGAGGTACCAAAAACAAGCCCTCCAAACAATGATGACCAGTGGAAAAACAATG
```

# ...Biological databases...

19/10/2022

- ...Introduction to Biological DB
- Bioinformatics centres of excellences
- Searching the database of interest

## Exploring some databases

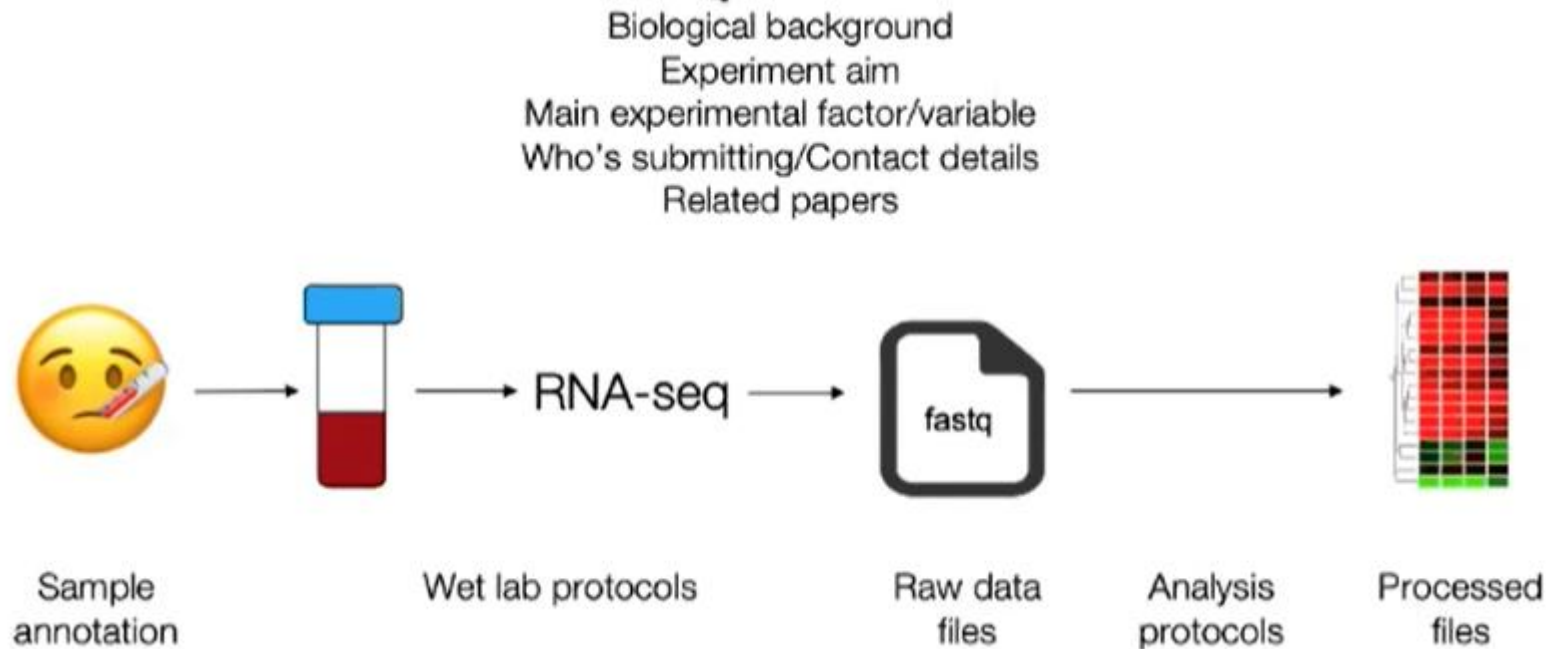
- ❖ Nucleotide Sequence (GeneBank, NCBI)
- ❖ Genetic variants (dbSNP, Clinvar, GnomAD, COSMIC)
- ❖ Protein Sequence (Uniprot, EMBL-EBI)
- ❖ Protei Interaction (Intact, EMBL-EBI)
- ❖ Gene-Phenotype associations (OMIM, HPO)
- ❖ Gene Expression (Gtex, Human Protein Atlas)
- ❖ Protein Expression (Human Protein Atlas)
- ❖ Human Metabolome (HMDB)



# What make a good bioinformatics DB: Primary vs derived data

	<b>Primary database</b>	<b>Secondary database</b>
Synonyms	Archival database	Curated database; knowledgebase
Source of data	Direct submission of experimentally-derived data from researchers	Results of analysis, literature research and interpretation, often of data in primary databases
Examples	<b>ENA, GenBank and DDBJ</b> (nucleotide sequence) <b>ArrayExpress</b> and <b>GEO</b> (functional genomics data) Protein Data Bank (PDB; coordinates of three-dimensional macromolecular structures)	InterPro (protein families, motifs and domains) UniProt Knowledgebase (sequence and functional information on proteins) Ensembl (variation, function, regulation and more layered onto whole genome sequences)

# What make a good bioinformatics DB: metadata



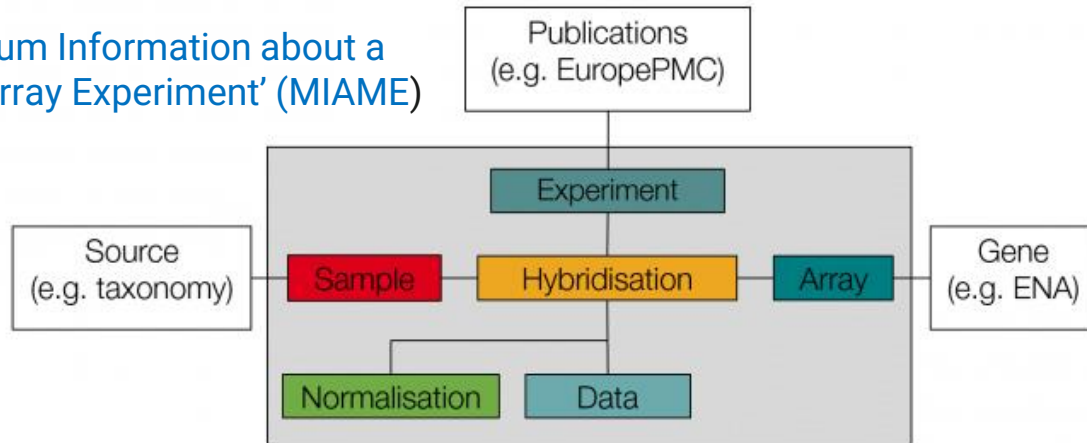
...the way in which biological data are recorded.

# What make a good bioinformatics DB: metadata standard

[Minimum information standards](#) are sets of guidelines and formats for reporting data derived by specific high-throughput methods

- data can be easily verified, analysed and interpreted by community
- facilitate the transfer of data from journal articles into databases
- available for a vast variety of experiment types

'Minimum Information about a Microarray Experiment' (MIAME)



...describing data consistently



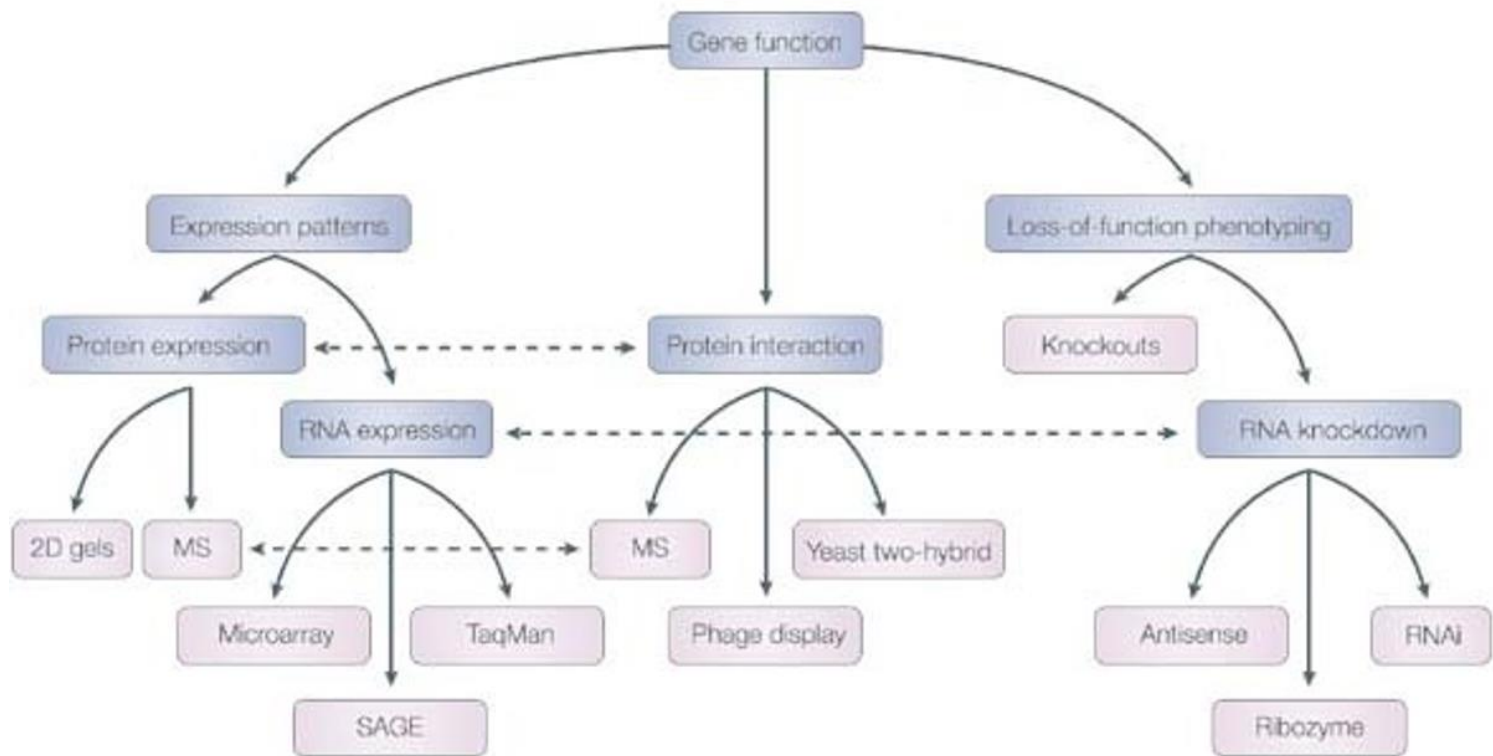
[ps://fairsharing.org/collection/MIBBI](https://fairsharing.org/collection/MIBBI)

# What make a good bioinformatics DB: controlled vocabulary

define specific words to reduce ambiguity and duplication

- non-hierarchical lists of terms
- use taxonomy as a classification scheme
- structured vocabulary in which concepts are represented by terms
- Using ontologies

# Heterogeneity in data (Scientific data domains)



# Some example of biological databases...

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb,  
ARR, AsDb, BBDB, BCGD, Beanref, Biolmage,  
BioMagResBank, BIOMDB, BLOCKS, BovGBASE,  
BOVMAP, BSORF, BTKbase, CANSITE, CarbBank,  
CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP,  
ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG,  
CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb,  
Picty\_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC,  
ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db,  
ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView,  
GCRDB, GDB, GENATLAS, Genbank, GeneCards,  
Genline, GenLink, GENOTK, GenProTEC, GIFTS,  
GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB,  
HAEMB, HAMSTERS, HEART-2DPAGE, HEXadb, HGMD,  
HIDB, HIDC, HlVdb, HotMolecBase, HOVERGEN, HPDB,  
HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat,  
KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB,  
Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5  
Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us,  
MPDB, MRR, MutBase, MycDB, NDB, NRSUB, O-lycBase,  
OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB,  
PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD,  
PFDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE,  
PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE,  
SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase,  
SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D,  
SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-  
MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB,  
TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE,  
VDRR, VectorDB, WDCM, WIT, WormPep, YEPD, YPD,  
YPM, etc ..... !!!!

# Where do I get DB of my interest ?

*D682–D688 Nucleic Acids Research, 2020, Vol. 48 Database issue*  
doi: 10.1093/nar/gkz966

Published online 6 November 2019

## Ensembl 2020

Andrew D. Yates , Premanand Achuthan, Wasiru Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, Andrey G. Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, José Carlos Marugán, Carla Cummins, Claire Davidson, Kamalkumar Dodiya, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, Thomas Maurel, Mark McDowall, Aoife McMahon, Shamika Mohanan, Benjamin Moore, Michael Nuhn, Denye N. Oheh, Anne Parker, Andrew Parton, Mateus Patricio, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M. Schmitt, Helen Schuilenburg, Dan Sheppard, Mira Sycheva, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Bethany Flint, Adam Frankish, Sarah E. Hunt, Garth Ilesley, Myrto Kostadima, Nick Langridge, Jane E. Loveland , Fergal J. Martin, Joannella Morales, Jonathan M. Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, Stephen J. Trevanion, Fiona Cunningham, Kevin L. Howe , Daniel R. Zerbino and Paul Flicek 

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 23, 2019; Revised October 09, 2019; Editorial Decision October 10, 2019; Accepted October 10, 2019

### ABSTRACT

The Ensembl (<https://www.ensembl.org>) is a system for generating and distributing genome annotation such as genes, variation, regulation and comparative

platform and programmatic interfaces (available under an Apache 2.0 license) and data updates made available four times a year.

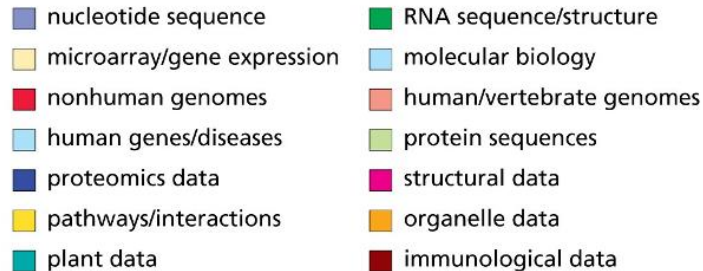
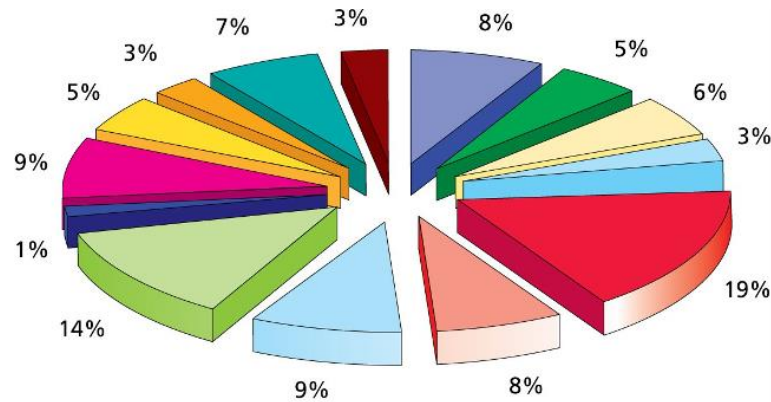
Downloaded from <https://academic.oup.com/nar/article/48/11/D682/5613662> by



# Nucleic Acid Research (NAR) Database Issue

Online collection of biological databases:

<http://www.oxfordjournals.org/nar/database/c/>



The list is not exhaustive!!

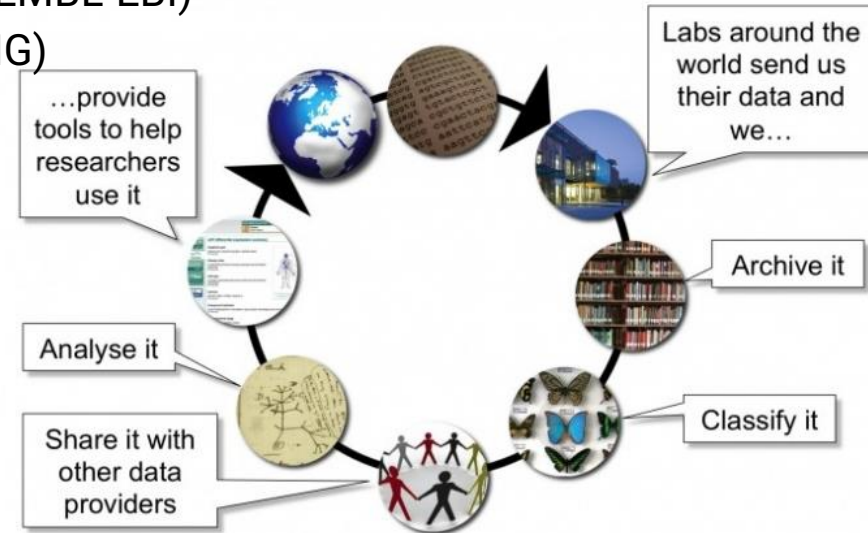


# Bioinformatics centres of excellences

*In the early 1980s DNA sequence data began to accumulate in the scientific literature.....*

Bioinformatics centres of excellences that collect, catalogue and provide open access to published biological data:

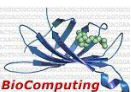
1. The US [National Center for Biotechnology Information](#) (NCBI)
2. The [EMBL-European Bioinformatics Institute](#) (EMBL-EBI)
3. [The National Institute of Genetics](#) in Japan (NIG)



# National Center for Biotechnology Information (NCBI)

The screenshot shows the NCBI website homepage. At the top, there is a blue navigation bar with the NIH logo, 'U.S. National Library of Medicine', and 'NCBI National Center for Biotechnology Information'. A 'Sign in to NCBI' link is on the right. Below this is a secondary navigation bar with links for 'NCBI HOME', 'LITERATURE', 'HEALTH', 'GENOMES', 'GENES', 'PROTEINS', 'CHEMICALS', and 'POPULAR RESOURCES'. A search bar is located below the navigation, with a dropdown menu currently set to 'All Databases' (indicated by a red arrow) and a 'Search' button. The main content area is titled 'About NCBI' and features a large image of a modern glass building. To the right of the image is a 'Follow Us' section with social media icons for Facebook, Twitter, Google+, LinkedIn, YouTube, RSS, Email, and a chat icon. Below that is an 'NCBI News & Blog' section with three news items: 'New version of PGAP now available!' (dated 13 Oct 2022), 'Now Available! Updated NCBI Datasets Command-Line Tools' (dated 12 Oct 2022), and 'Now available: Updated prokaryote representative genomes collection' (dated 11 Oct 2022). The 'About NCBI' section is divided into four columns: 'Our Mission' (describing the goal of uncovering new knowledge), 'Organizational Structure' (describing the role of branches and the Board of Scientific Counselors), 'Programs & Activities' (describing resources for genomic, genetic, and biomedical data), and 'Researchers at NCBI' (describing the basic research program conducted by intramural investigators). Each column has a small icon representing its theme.

<https://www.ncbi.nlm.nih.gov/home/about/>



# National Center for Biotechnology Information (NCBI)

## Health



NCBI's Health resources include databases for use in clinical practice and medical research that contain information about human disease and pathology, including diagnostics and treatments.

### How to

- Find genes associated with a condition
- Find variations with a clinical assertion for a condition
- View genotype frequency for a gene or condition
- Find a clinical practice guideline for a condition

more...

### Clinical & Public Health Resources

- MedGen**  
human medical genetics
- Genetic Testing Repository (GTR)**  
genetic test information and laboratories
- clinical effectiveness, disease and drug reports
- ClinicalTrials.gov**  
clinical trials registry and study results
- Pathogen Detection Project**  
microbial genome sequence analysis for epidemiologic surveillance

### Human Variation

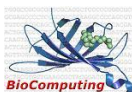
- ClinVar**  
human variations of clinical significance
- RefSeqGene**  
reference standard human genome sequences
- dbGaP**  
genotype/phenotype interaction studies
- OMIM**  
Online Mendelian Inheritance in Man database

### Literature

- PubMed**  
scientific and medical citations
- PubMed Clinical Queries**  
preformatted clinically-based PubMed queries
- GeneReviews**  
genetic disease reviews on the Bookshelf

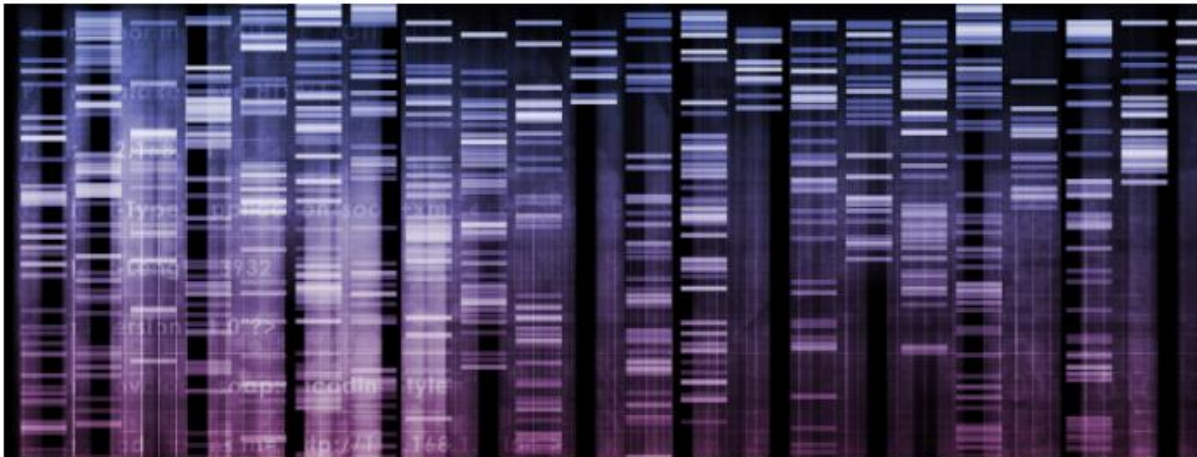
## Documentation

Online manuals, handbooks, fact sheets and FAQs



# National Center for Biotechnology Information (NCBI)

## Genes



### Gene Loci

#### Gene

gene summary information and links

clusters of expressed sequences

#### Nucleotide

gene and transcript sequences

### Homologs

#### HomoloGene

homologous gene sets for selected organisms

#### PopSet

studies of sequences within and across populations

#### Protein Clusters

sequence similarity-based protein

### Gene Expression

#### GEO Profiles

expression profiles for individual genes


#### EST

expressed sequence tags

#### SRA

high-throughput DNA sequences

# NCBI: GEO profiles

 **National Library of Medicine**  
National Center for Biotechnology Information

GEO Profiles

Summary ▾

[ATP1A2 - Rett syndrome: brain frontal cortex](#)

Annotation: ATP1A2, ATPase Na+/K+ transporting subunit alpha 2

Organism: Homo sapiens

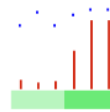
Reporter: GPL8300, 34377\_at (ID\_REF), GDS2613, 477 (Gene ID), J05096

DataSet type: Expression profiling by array, count, 6 samples

ID: 36029718

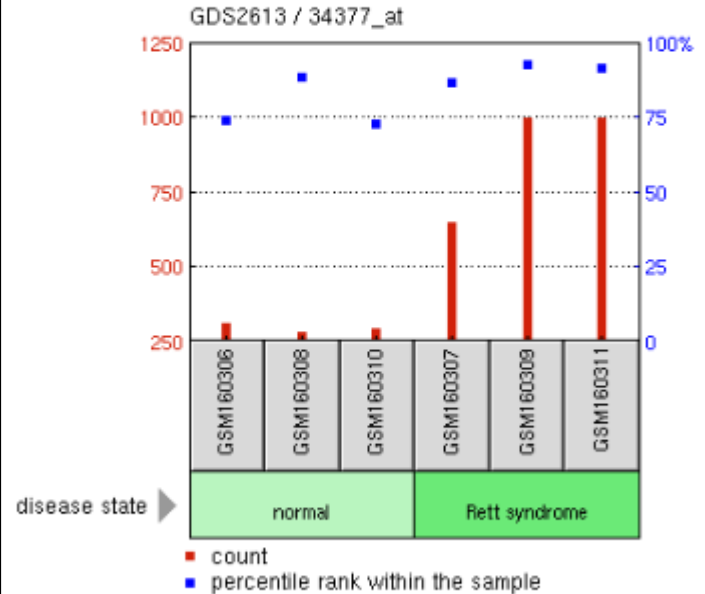
[GEO DataSets](#) [Gene](#) [Profile neighbors](#) [Chromosome neighbors](#) [Homologene neighbors](#)

Click the graph



This database stores individual gene expression profiles from curated DataSets in the Gene Expression Omnibus (GEO) repository.

**Profile** GDS2613 / 34377\_at  
**Title** Rett syndrome: brain frontal cortex  
**Organism** Homo sapiens



[Graph caption help](#)



# EMBL's European Bioinformatics Institute EMBL-EBI

The EMBL-EBI website has been redesigned. Please send us feedback about this page.

EMBL's European Bioinformatics Institute

## EMBL-EBI

Unleashing the potential of big data in biology



Example searches: blast keratin bft1 | About EBI Search

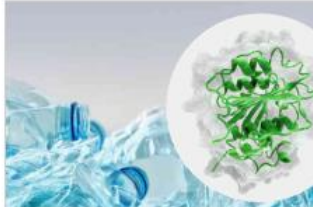
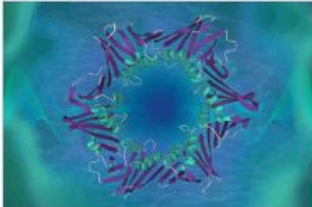
[Find data resources](#) →

[Submit data](#) →

[Explore our research](#) →

[Train with us](#) →

[Latest news](#) →



# EMBL's European Bioinformatics Institute EMBL-EBI

## EMBL-EBI data resources and tools

EMBL's European Bioinformatics Institute maintains the world's most comprehensive range of freely available and up-to-date molecular data resources.



Find a data resource or a tool

Showing 1 - 15 out of 156 results




Sort by:

Refine by  
Type

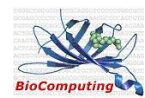
- Data resources
- Tools

Category

- Chemical biology
- Cross domain
- DNA & RNA
- Gene expression
- Literature
- Ontologies
- Proteins

-  **AlphaFold DB**  
Database for protein structure predictions for numerous species  
[DATA RESOURCE](#) [CC-BY](#)
-  **BioModels**  
A repository of peer-reviewed, published, computational models.  
[DATA RESOURCE](#) [Web API](#) | [CCO](#)
-  **ChEMBL**  
An open data resource of binding, functional and ADMET bioactivity data.  
[DATA RESOURCE](#) [Web API](#) | [CC-BY](#)

[AlphaFold](#) is an AI system developed by [DeepMind](#) that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.



# EMBL's European Bioinformatics Institute EMBL-EBI

EMBL-EBI

Services

Research

Training

About us



EMBL-EBI Logo

Login Register

## EMBL-EBI Training

Delivering world-class training in data-driven life sciences.

Coronavirus information: the delivery format of EMBL-EBI events is being continually evaluated. Please check the individual event listing to see if the course will be run in-person or virtually. Your safety is paramount to us; you can read our [COVID guidance policy](#) for more information.

Search

### Featured courses



LIVE WEBINAR

A guide to UniProt for students

Open | 🗓️ 27 October 2022 | 📍 Online



COURSE AT EMBL-EBI

Introduction to RNA-seq and functional interpretation

Applications close: 6 November 2022 | 🗓️ 20 - 24 February 2023 | 📍 European Bioinformatics Institute, United Kingdom



VIRTUAL COURSE

Single-cell RNA-seq analysis using Galaxy

Applications close: 27 November 2022 | 🗓️ 6 - 10 February 2023 | 📍 Online

From EMBL-EBI we will see  
UniProt and Intact





# Database of Protein Sequence

## Contact the UniProt consortium members

EMBL-EBI



EMBL Outstation  
European Bioinformatics Institute (EBI)  
Wellcome Trust Genome Campus  
Hinxton Cambridge CB10 1SD  
United Kingdom  
Phone: (+44 1223) 494 444  
Fax: (+44 1223) 494 468



SIB Swiss Institute of Bioinformatics  
Centre Medical Universitaire  
1, rue Michel Servet  
1211 Geneva 4  
Switzerland  
Phone: (+41 22) 702 50 50  
Fax: (+41 22) 702 58 58



Protein Information Resource (PIR)  
Georgetown University Medical Center  
3300 Whitehaven Street NW  
Suite 1200  
Washington, DC 20007  
United States of America  
Phone: (+1 202) 687 1039  
Fax: (+1 202) 687 0057

# Uniprot

<http://www.uniprot.org/>



# Database of protein sequences - Uniprot

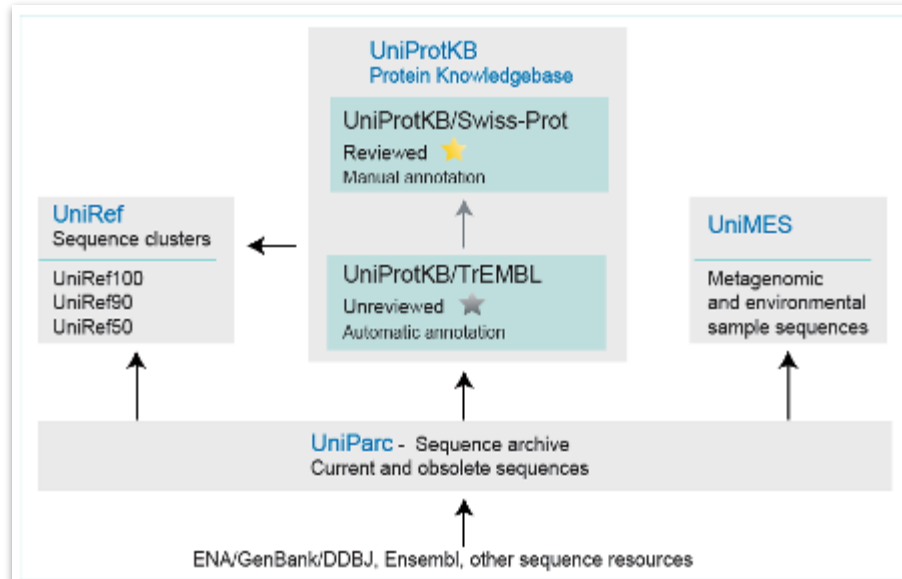
**Uniprot knowledgebase** (UniprotKB) consists of two sections:



- **Swiss-Prot**, which is manually annotated and reviewed

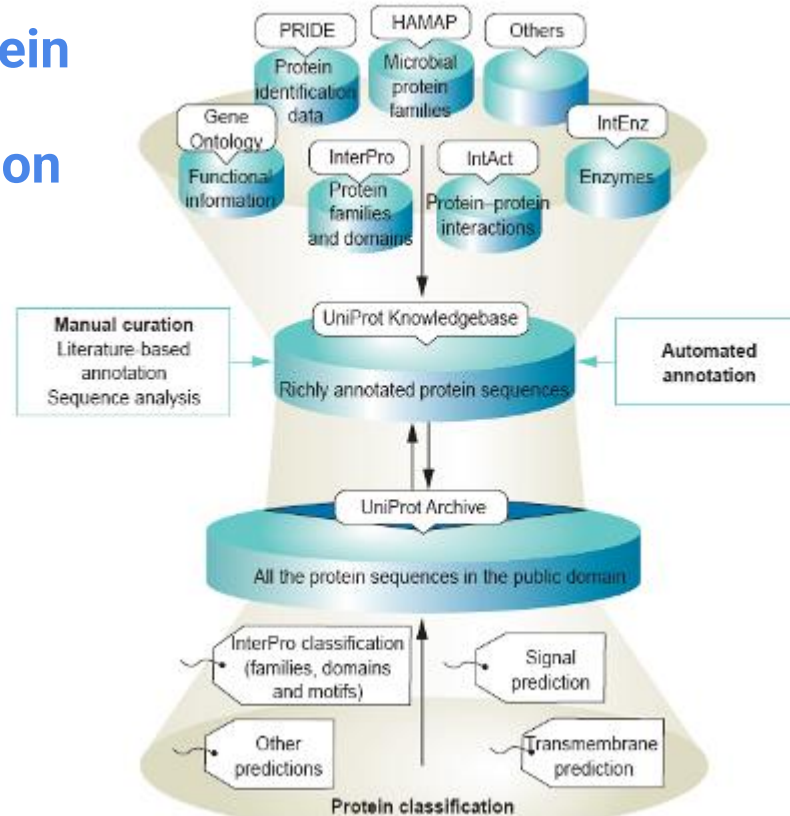


- **TrEMBL**, which is automatically annotated and is **not** reviewed

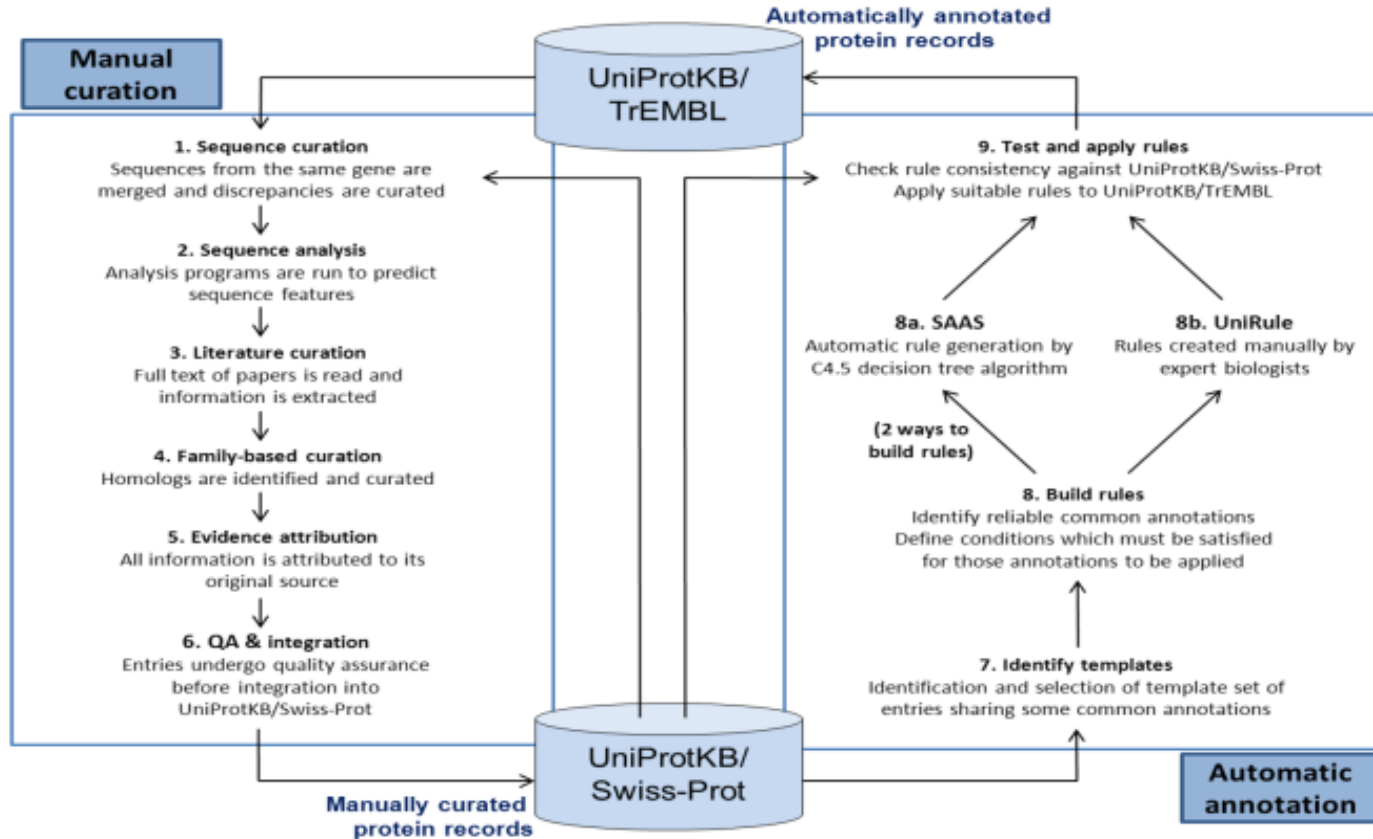


# Database of protein sequences - Uniprot

Is a resource of **protein sequences** and **functional information**



# Database of protein sequences - Uniprot



# IntAct Molecular Interaction Database

IntAct provides a free, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions. The IntAct Team also produces the [Complex Portal](#). You are currently visiting the new website of IntAct. The former version can be found [here](#) and will be supported until the end of 2021.

 Newsletter

email address

## Datasets

Datasets

### Datasets of biological significance

We provide 4 types of interaction datasets:

#### Topical:

Manually curated datasets that are either [manually](#) or [computationally](#) assigned to a specific biological topic.

#### Interactomes:

e.g. Rare disease, Neurodegeneration, Cancer

For 16 different [species](#).

#### Mutations:

Annotations of experimental evidence where [mutations](#) have been shown to affect a molecular interaction.

## EMBL-EBI resource

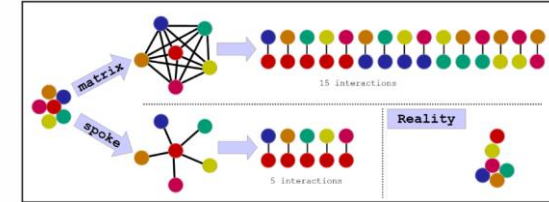
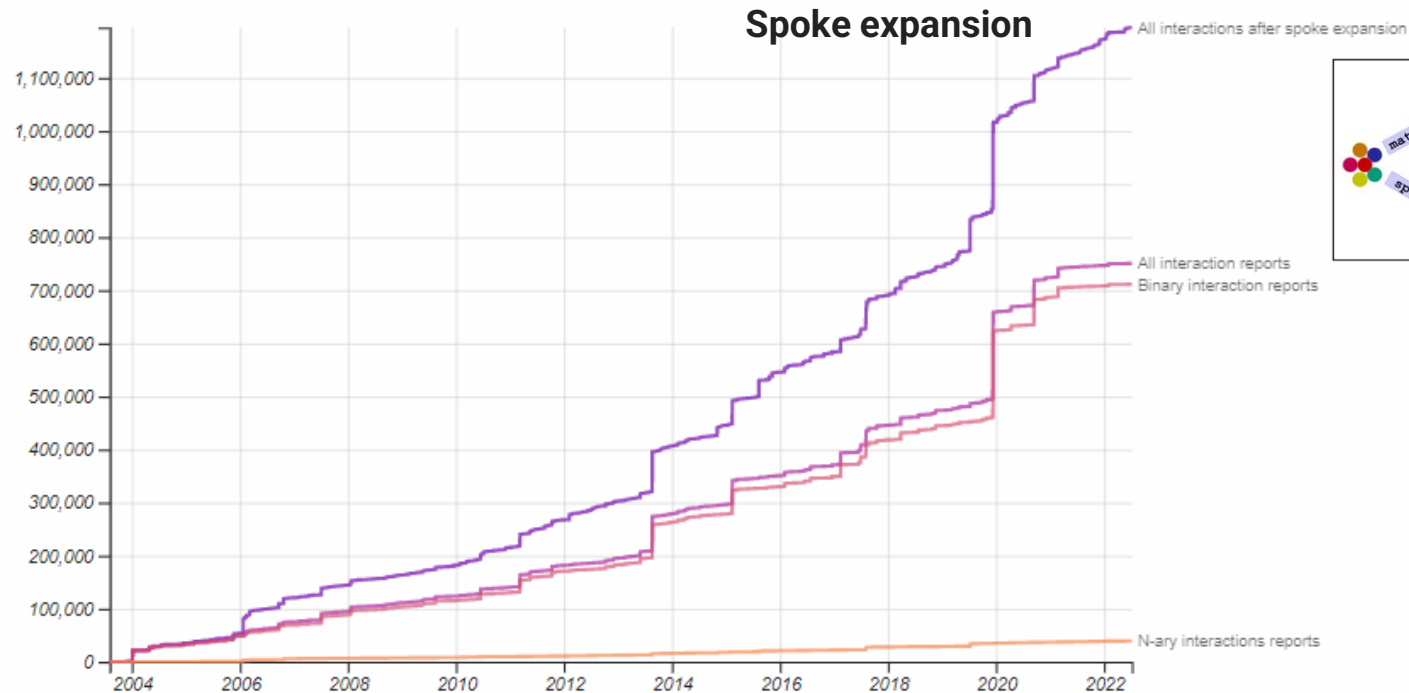
<https://www.ebi.ac.uk/intact/home>

# EMBL-EBI resource: IntAct

## Statistics

The details below are based on released content of the IntAct database, with contributions from all IMEx partners. Move your cursor over the graphs for further details.

## Interactions over time

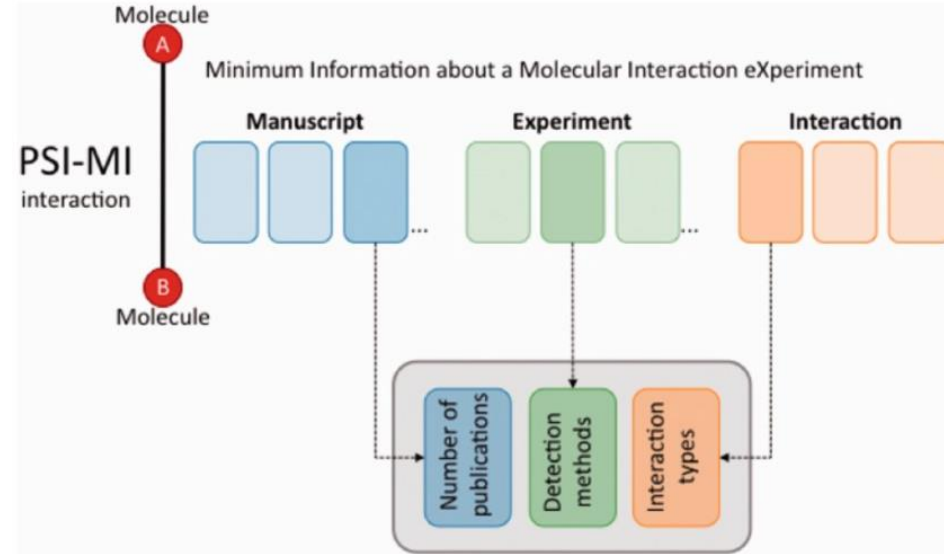


# EMBL-EBI resource: IntAct

## Interaction Scoring MIscore

Customizable, heuristic scoring system that takes three factors into account:

1. How the interaction was observed, predicted or inferred (interaction detection method; MI:0001)
2. The type of interaction. Direct interaction, physical association, co-localization and so forth. (interaction type; MI:0190)
3. The number of publications reporting a specific interaction



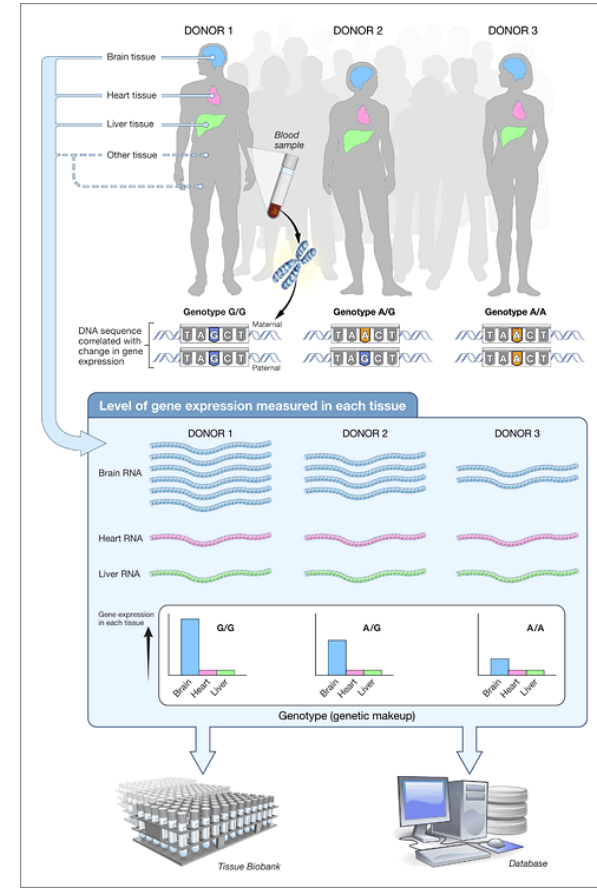
# Genotype-Tissue Expression GTEx Portal

(<https://www.gtexportal.org/home/>)

The screenshot shows the GTEx Portal homepage. At the top left is the GTEx Portal logo. To the right are navigation links: About GTEx, Publications, and Access Biospecimens. Below this is a dark navigation bar with links for Home, Downloads, Expression, Single Cell, QTL, IGV Browser, Tissues & Histology, and Documentation, along with a search bar for genes or SNPs. A pink banner below the navigation bar contains a survey link. The main content area features a large image with a DNA double helix and a text box titled '2019-08-28 GTEx Portal V8 Release' which states that the V8 release includes 17,382 RNA-Seq samples from 948 donors, representing an increase of 49% and 33% relative to V7. Below the image are two buttons: 'Resource Overview' and 'Explore GTEx'.







The GTEx Project ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. 54 non-diseased tissue sites across nearly 1000 individuals (with WGS, WES, and RNA-Seq)

The Developmental dGTEx Project is a new effort to study developmental-specific genetic effects on gene expression.





# Genotype-Tissue Expression GTEx Portal

Explore GTEx					
 Browse	<input type="text" value="By gene ID"/>	Browse and search all data by gene	 QTL		
	<input type="text" value="By variant or rs ID"/>	Browse and search all data by variant		<a href="#">Locus Browser (Gene-centric)</a>	Visualize QTLs by gene in the Locus Browser
	<a href="#">By Tissue</a>	Browse and search all data by tissue		<a href="#">Locus Browser (Variant-centric)</a>	Visualize QTLs by variant in the Locus Browser VC (Variant Centric)
	<a href="#">Histology Viewer</a>	Browse and search GTEx histology images		<a href="#">IGV Browser</a>	Visualize tissue-specific eQTLs and coverage data in the IGV Browser
 Single Cell	<a href="#">Data Overview</a>	Learn more about available single cell data	<a href="#">eQTL Dashboard</a>	Batch query eQTLs by gene and tissue	
	<a href="#">Multi-Gene Single Cell Query</a>	Browse and search single cell expression by gene and tissue		<a href="#">eQTL Calculator</a>	Test your own eQTLs
 Expression	<a href="#">Multi-Gene Query</a>	Browse and search expression by gene and tissue	 eGTEX	<a href="#">H3K27ac, m6A, WGBS</a>	Browse H3K27ac ChIP-seq, m6A methylation, and WGBS DNA methylation data in IGV Browser
	<a href="#">Transcript Browser</a>	Visualize transcript expression and isoform structures			 Biobank

<https://www.proteinatlas.org/>

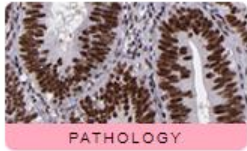
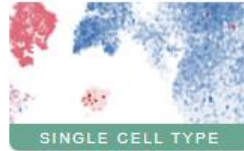
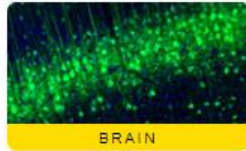
# THE HUMAN PROTEIN ATLAS

≡ MENU HELP NEWS

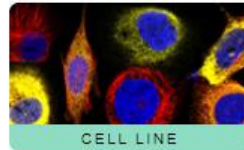
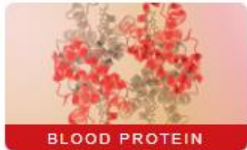
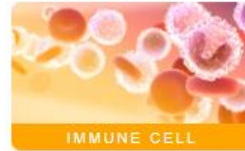
SEARCH

e.g. ACE2, GFAP, EGFR

Search Fields »



The open access resource for human proteins  
Search for specific genes/proteins or  
explore the 10 different sections



The Human Protein Atlas is a Swedish-based program initiated in 2003


**Aim:** to map all the human proteins in cells, tissues, and organs using an integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics, and systems biology.

# Human Metabolome Database

**HMDB** Browse Search Downloads About Contact Us

Search metabolites Search

**TMIC** The **Metabolomics Innovation Centre** Quantitative metabolomics services for biomarker discovery and validation.

  
**hmdb**  
The Human Metabolome Database

Browse Metabolites >>

Learn More >>

What's New >>

# Human Metabolome Database

The database is designed to contain or link three kinds of data:

1. chemical data
2. clinical data
3. molecular biology/biochemistry data.

220,945 metabolite entries including both water-soluble and lipid soluble metabolites.  
8,610 protein sequences (enzymes and transporters) are linked to these metabolite entries.

Many data fields are hyperlinked to other databases  
([KEGG](#), [PubChem](#), [MetaCyc](#), [ChEBI](#), [PDB](#), [UniProt](#), and [GenBank](#))

[DrugBank](#) contains equivalent information on ~2832 **drugs** and 800 drug metabolites  
[T3DB](#) contains information on ~3670 common **toxins** and environmental pollutants  
[SMPDB](#) contains pathway diagrams for ~132,335 human metabolic, drug and disease pathways

**Now we can start exploring the  
databases!**