# Building bioinformatics resources

Qingyao Huang | UZH BIO390 HS25

# Bioinformatics history

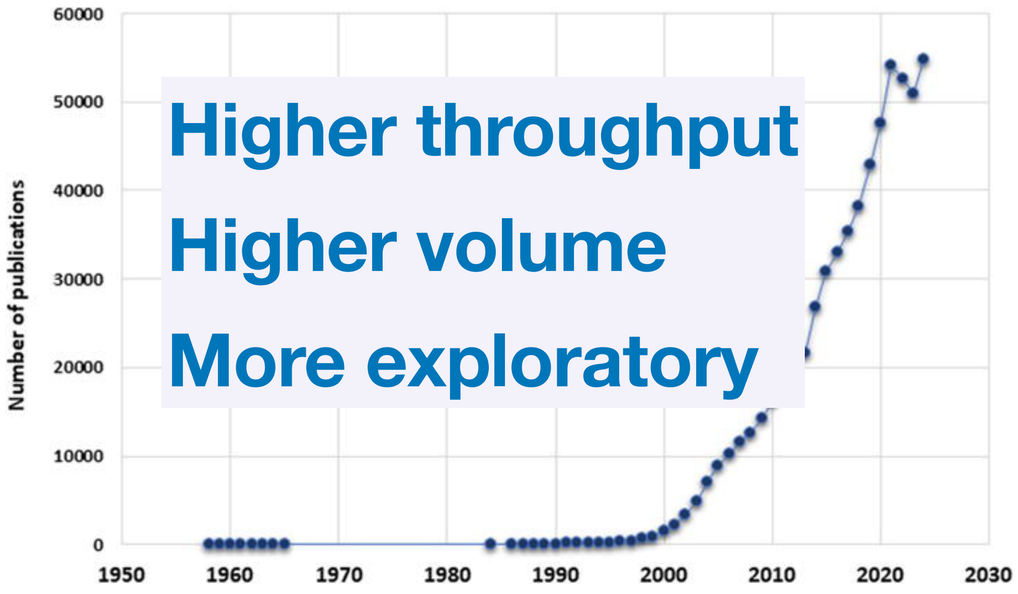
**1950s-1970s**
**DNA Structure** Discovery (1953)
DNA **Sequencing** (Sanger, 1950s)
Early Algorithms for protein folding (eg, Levinthal's Paradox)

**1980s-1990s**
Sequence Databases **GenBank** (1982), **PDB** (3D structure data)
**Pairwise Alignment**. Needleman-Wunsch (1970), Smith-Waterman (1981)
BLAST (1990): Fast sequence searching
**First Genome** Sequenced (H. influenzae, 1995)

**2000s (Genomics Revolution)**
**Human Genome** Project (Completed 2003)
**Microarray** Technology: Gene expression profiling
Next-Generation Sequencing (**NGS**): Illumina, 454 (Mid-2000s)
ENCODE Project (Started 2003): **Functional annotation** of the genome

**2010s (Systems Biology & Big Data)**
**Multi-omics**: Genomics, transcriptomics, proteomics, metabolomics
**Gene editing** (CRISPR-Cas9: guide RNA design)
**Single-Cell** RNA-Seq / live-cell imaging
AI/ML in Bioinformatics **Deep learning**

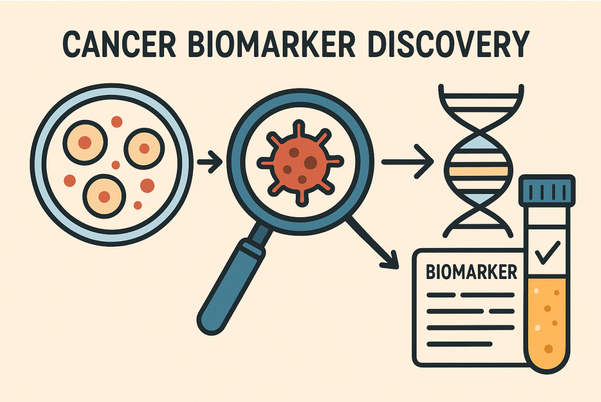
**2020s & Beyond**
**Long-Read** Sequencing (PacBio, Nanopore)
**Personalized** Medicine: Pharmacogenomics, cancer
AlphaFold & **Protein Structure Prediction** (2020)
**Metagenomics & Microbiome**




Abdurakhmonov and Ayubov Latest Trends in Bioinformatics Development 2025

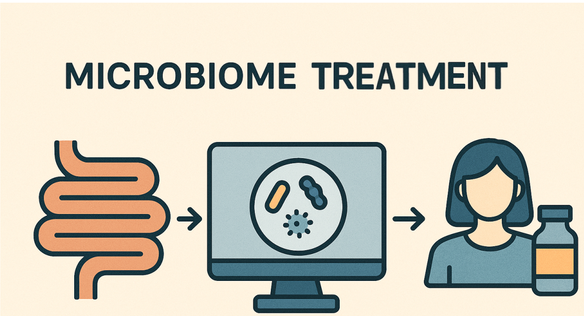
# Bioinformatics in healthcare, medicine

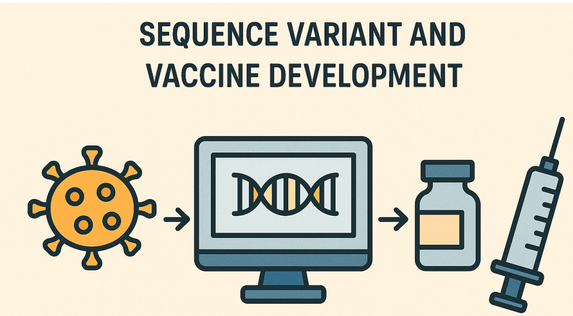
ChEMBL
PDB


RefSeq,
GTEx

GMrepo

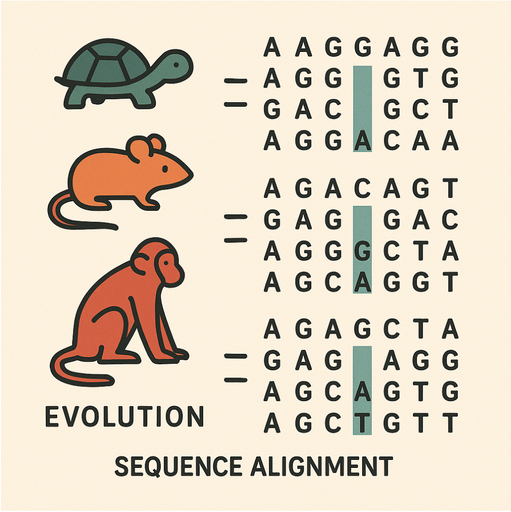


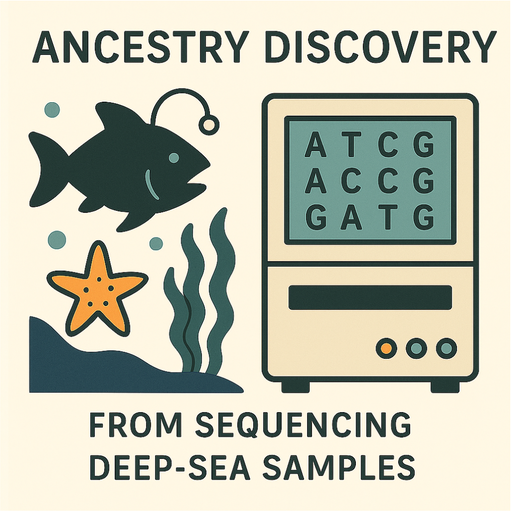




GISAID

# Bioinformatics in fundamental biology

- Explore the unknown
- Understand mechanisms

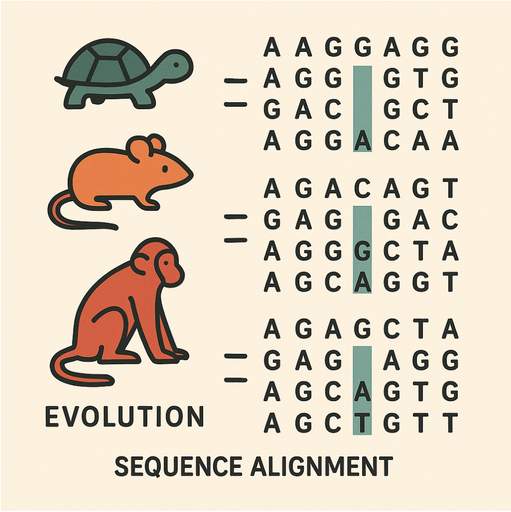




Sprang et al, Nature 2015

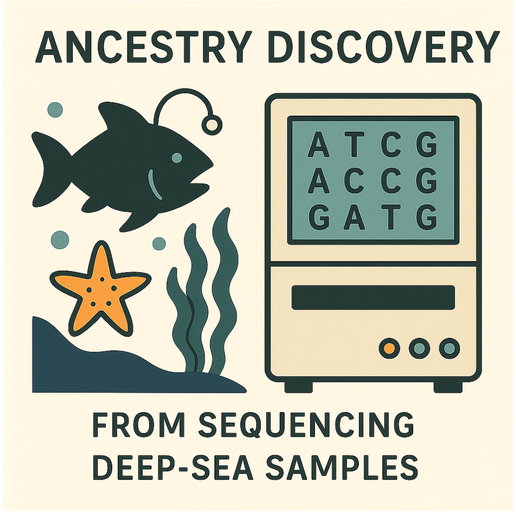
# Bioinformatics in fundamental biology

KEGG, Reactome


eggNOG, OMA




GTDB, NCBI Taxonomy

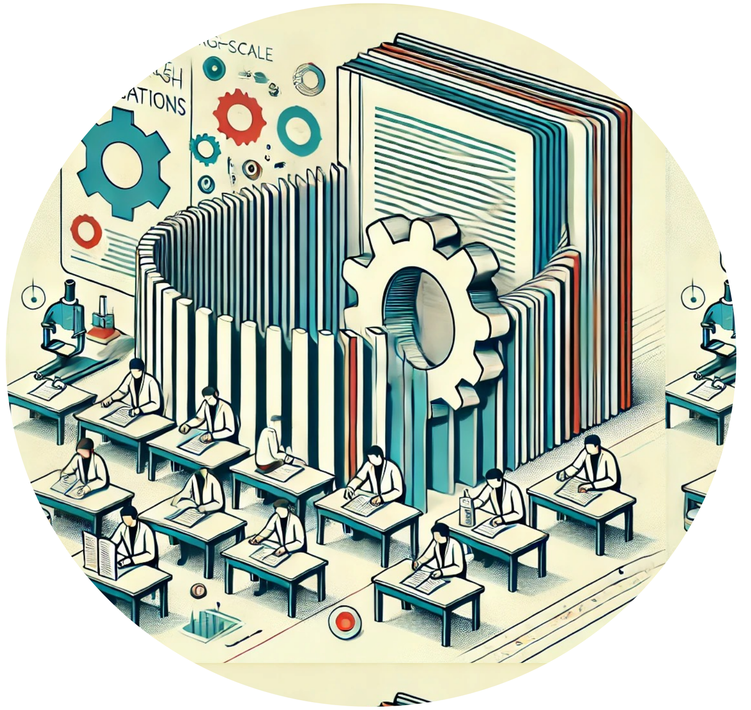



Sprang et al, Nature 2015

# Different classes of databases

- Primary data repository
- Curated database
- Meta-database / knowledge-bases

# Primary data repositories

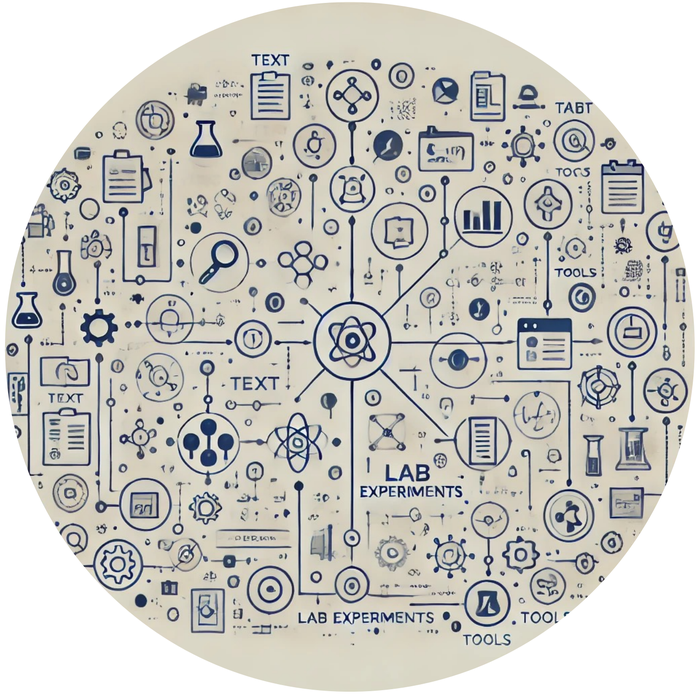


- Collects primary datasets conducted by individual groups
- Archival type, reproducibility
- Examples:
  - Gene Expression Omnibus (GEO; NCBI)
  - ArrayExpress (EBI)
  - GenBank
  - Protein Data Bank (PDB)
  - Proteomics Identification Database (PRIDE; EBI)

# Curated databases

- Codify terms, summarize knowledge
- Examples:
  - Functional pathways:
    - KEGG
    - Gene Ontology (GO)
    - Reactome
  - Ontologies:
    - Disease ontology
    - BRENDA tissue ontology
    - Cell Line Ontology …

# Meta-databases / knowledge-bases




- Integrates multiple data sources with statistics and provides a knowledge portal for a domain
- Examples:
  - UniprotKB
  - InterPro
  - Progenetix
  - STRING

# FAIR data principle

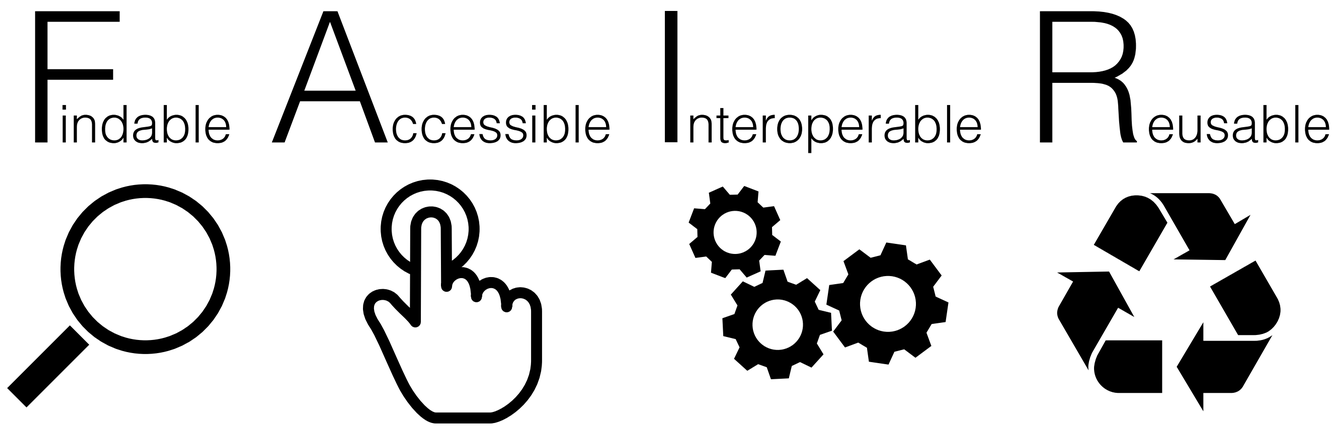


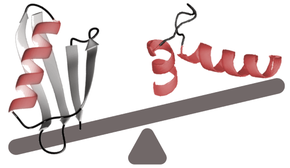

Image source: wikipedia

**Easier to use, manage and collaborate**

**→ increase value of the data**

# PaxDb

## A Protein abundance reference resource



paxdb

# PaxDb

- Motivation for building a DB
  - Relevance
- What is the quantitative data and how are data included?
  - Protein abundances
  - Techniques
- What are the metadata?
  - species, tissue, protein ID, ortholog
  - publication, experimental condition
- How to access the resources?
  - Web browsing, bulk download, upload own data

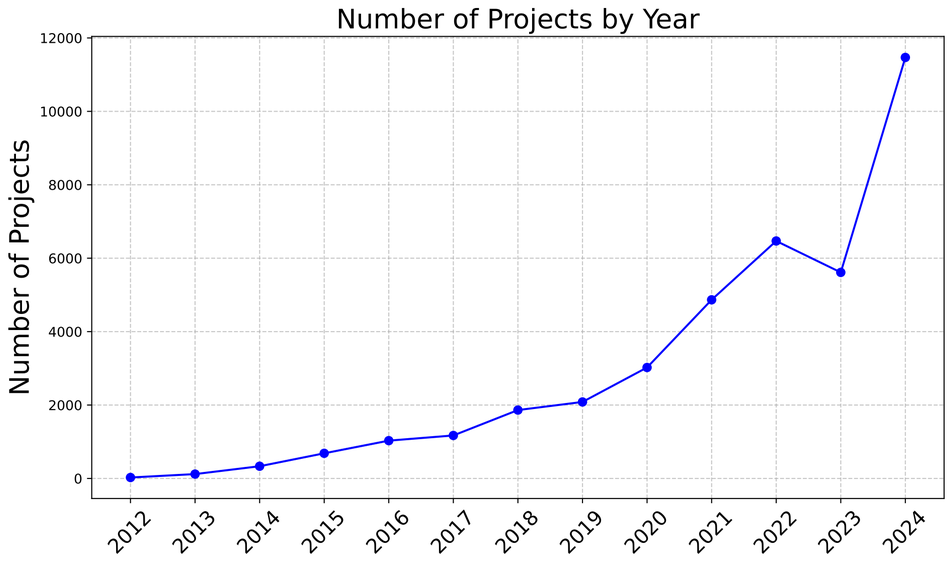
# PaxDb

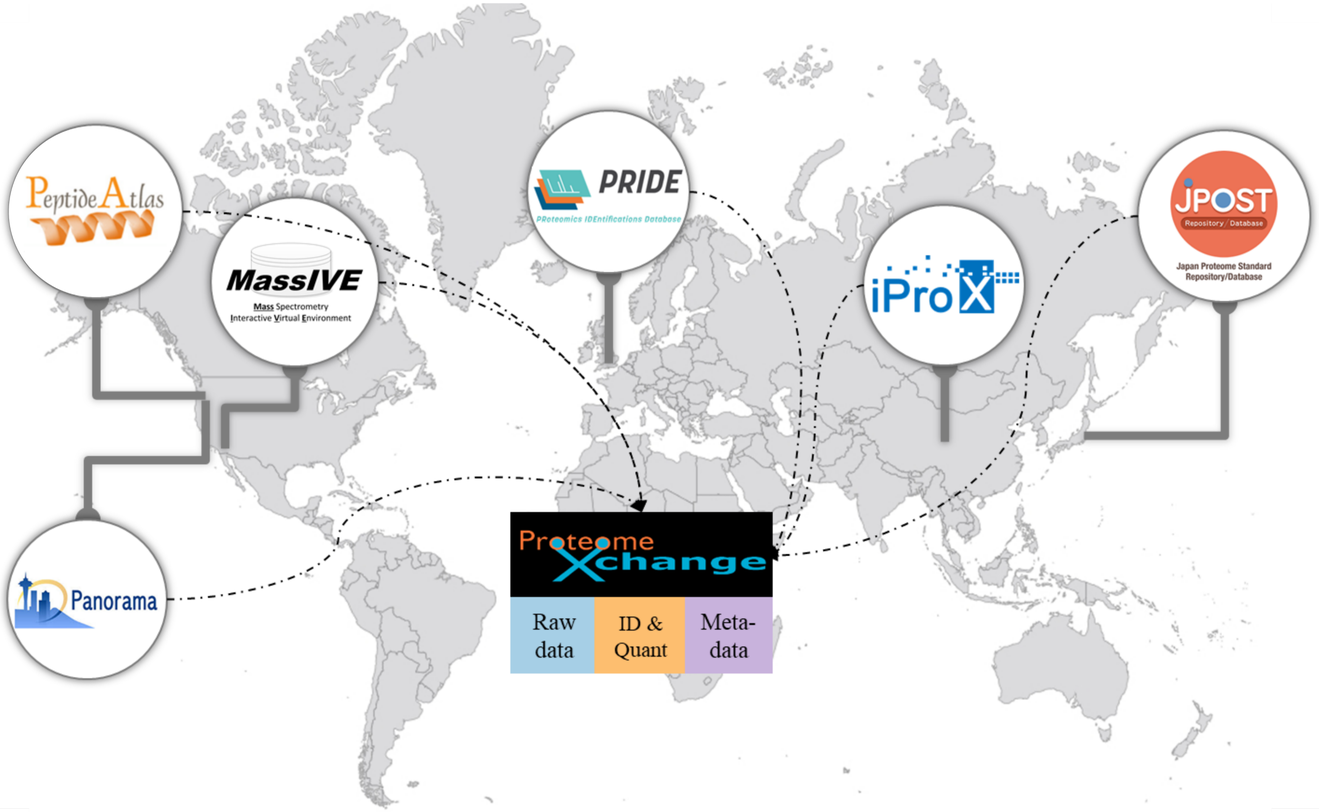
## Motivation

- Protein abundance across organisms
- Proteomics datasets are large and difficult to process and compare
- Reference for
  - human tissues
  - common and rare species
  - cross-species comparison
- Integrated on tissue / organism level

# ProteomeXchange

- Data registry (consortium) of multi-regional data repositories



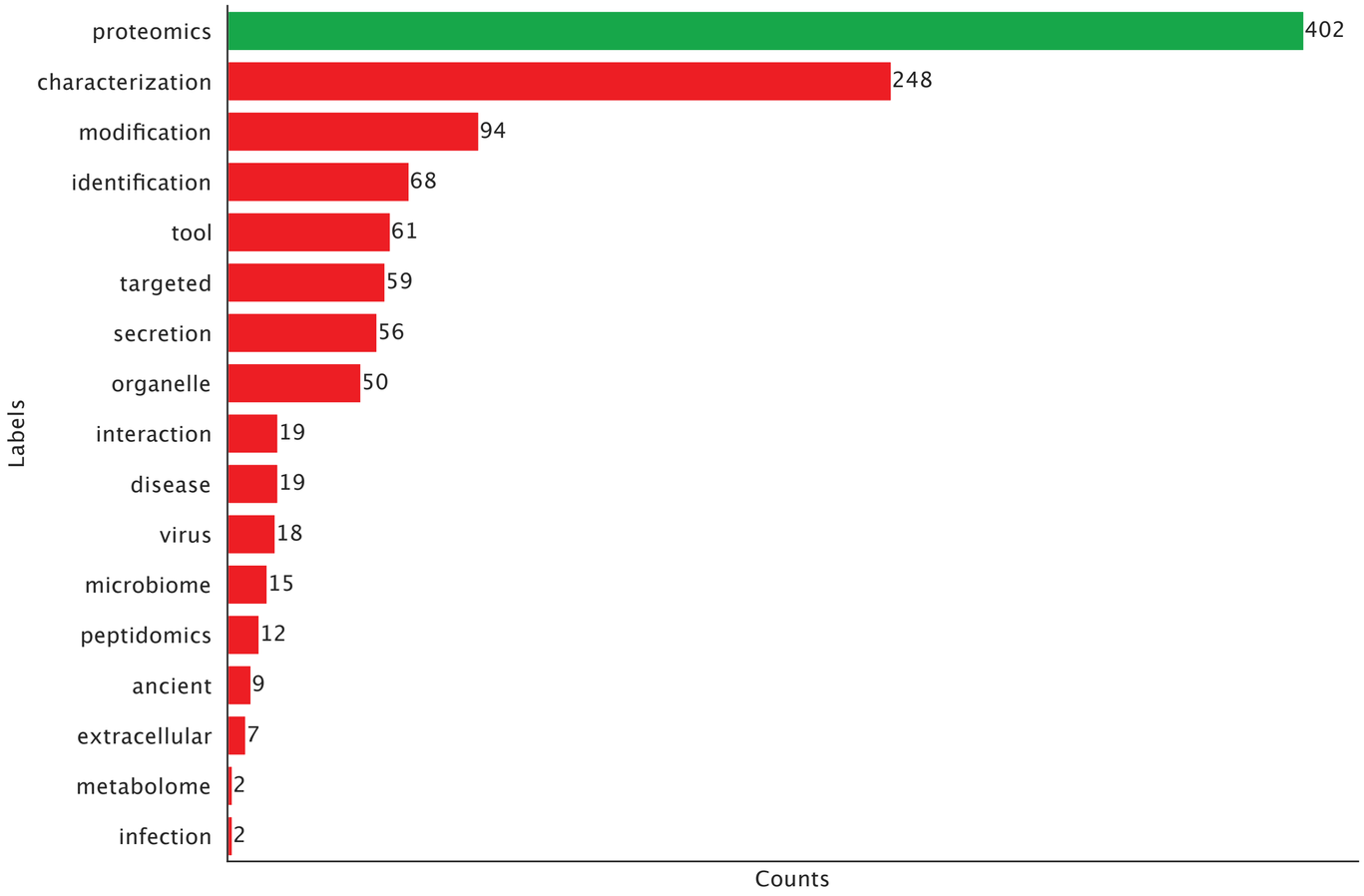
# Project selection


## Scope

Whole proteome
Only healthy/normal
Quantitative
No sub-cellular fraction
No mixed species
No PTM

# LLM-assisted Project selection

Training data: 1141 title + abstract




**Topic modeling (TM)**

**Embedding classification (EC)**

**GPT Chat completion API**

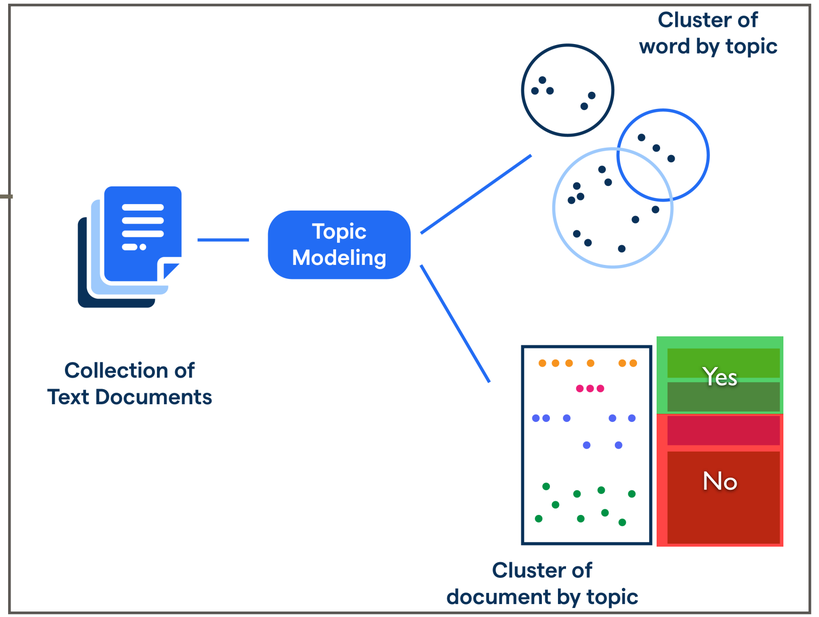
# LLM-assisted Project selection

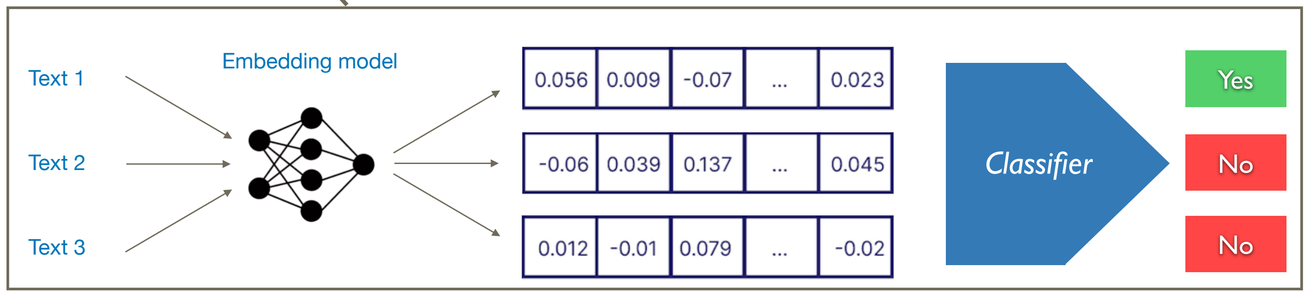
## Topic modeling (TM)

## Embedding classification (EC)

## GPT Chat completion API

You are a helpful assistant. Here are several studies with their titles and abstracts. Please evaluate each study to determine whether it focuses on quantitative proteomics with the specific criteria:
Inclusion criteria: ……
Exclusion criteria: ……
Title: ……
Abstract: ……




https://botpenguin.com/glossary/topic-modeling

# Method comparison

## Topic modeling (TM)

## Embedding classification (EC)

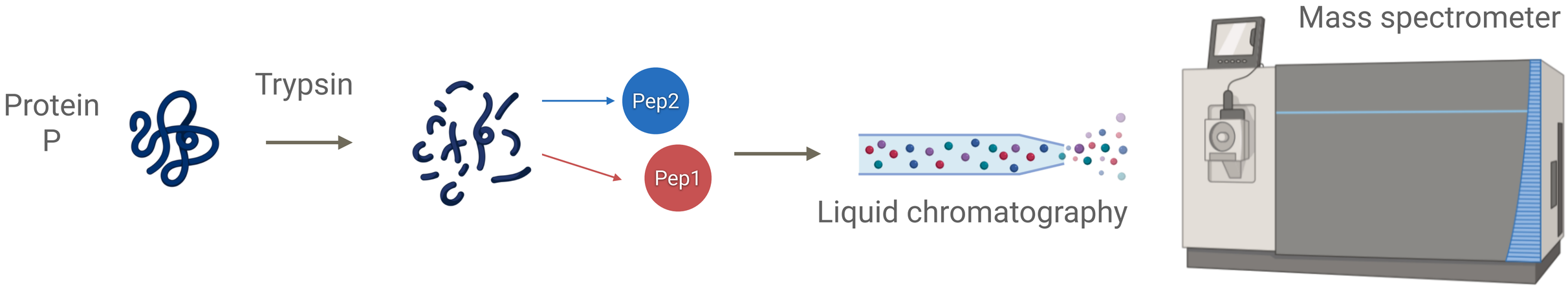
## GPT Chat completion API

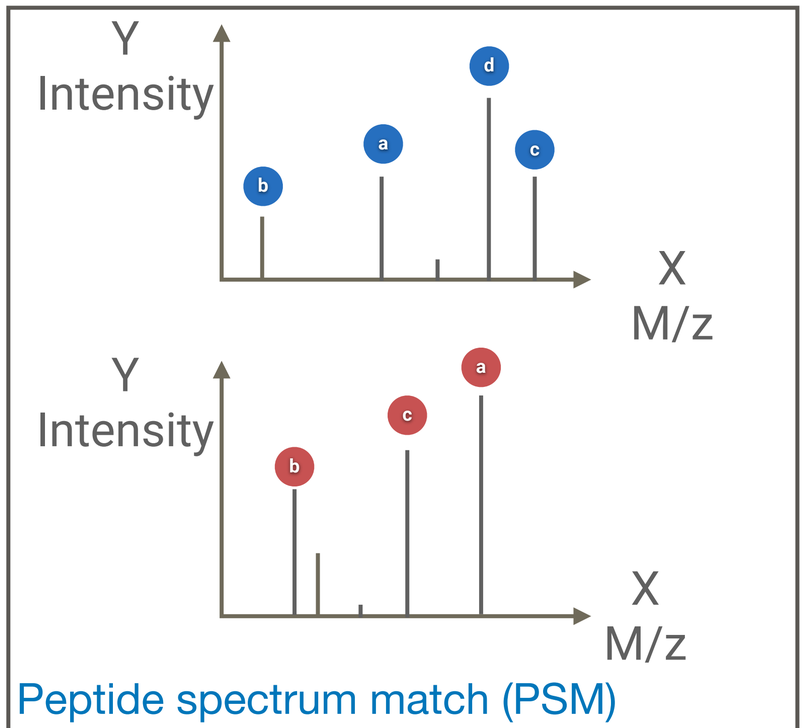
## Keyword (quant, abundance)

| Method | Instances | Recall | Precision | F1 |
|---|---|---|---|---|
| TM | R10_6_5_3 | 0.609 | 0.786 | 0.691 |
| | S10_6_5_2 | 0.804 | 0.62 | 0.682 |
| | R11_8_5_1 | 0.77 | 0.6 | 0.691 |
| EC | FT008 | 0.767 | 0.747 | 0.757 |
| | FT010 | 0.562 | 0.854 | 0.678 |
| | CM109 | 0.795 | 0.58 | 0.671 |
| OpenAI | GPT-3.5 | 0.798 | 0.515 | 0.626 |
| | GPT-4o | 0.612 | 0.711 | 0.658 |
| Keyword | | 0.672 | 0.512 | 0.581 |

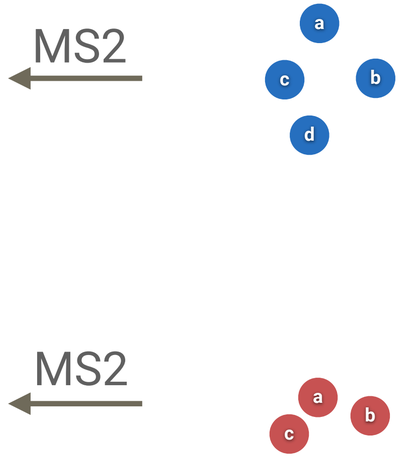
→ Majority voting from the best models
Integrated into continuous data processing

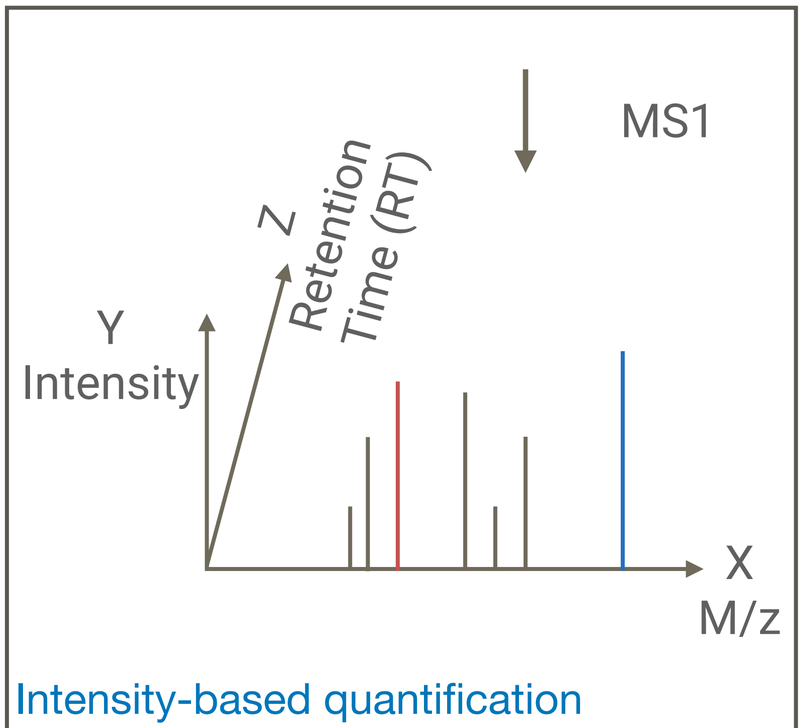
# Quantitative Proteomics with LC-MS/MS




Peptide spectrum match (PSM)

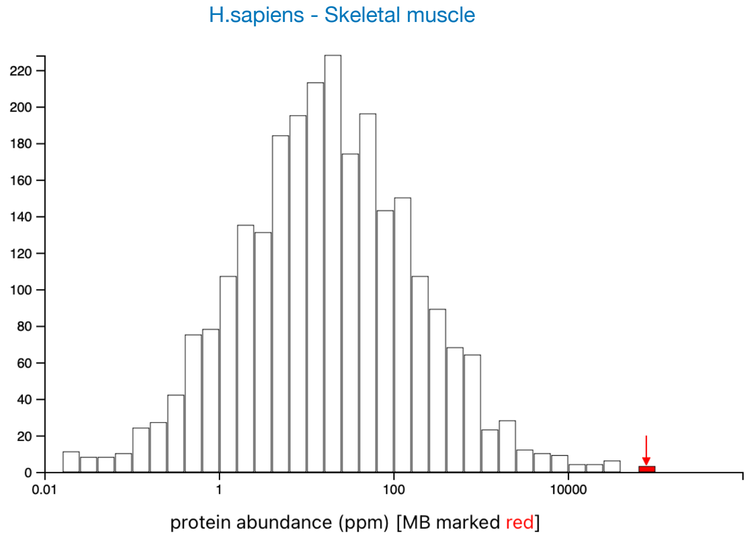



Intensity-based quantification

# From MS1 intensity to protein quantification

- Match peptide sequence to reference proteome
- Protein abundance = sum of peptides' intensity normalized by protein size
- normalize against total proteome abundance




| | Protein | Abundance |
|---|---|---|
| 1 | MB | 95525 |
| 2 | CA3 | 81288 |
| 3 | CKM | 71163 |
| 4 | ALB | 37135 |
| 5 | FHL1 | 36272 |

>9606.ENSP00000001008
MTAEEMKATESGAQSAPLPMEGVDISPKQDEGVLKVIKRE
GTGTEMPMIGDRVFVHYTGWLLDGTKFDSSLDRKDKFSFD
LGKGEVIKAWDIAIATMKVGEVCHITCKPEYAYGSAGSPP
KIPPNATLVFEVELFEFKGEDLTEEEDGGIIRRIQTRGEG
YAKPNEGAIVEVALEGYYKDKLFDQRELRFEIGEGENLDL
PYGLERAIQRMEKGEHSIVYLKPSYAFGSVGKEKFQIPPN
AELKYELHLKSFEKAKESWEMNSEEKLEQSTIVKERGTVY
FKEGKYKQALLQYKKIVSWLEYESSFSNEEAQKAQALRLA
SHLNLAMCHLKLQAFSAAIESCNKALELDSNNEKGLFRRG
EAHLAVNDFELARADFQKVLQLYPNNKAAKTQLAVCQQRI
RRQLAREKKLYANMFERLAEEENKAKAEASSGDHPTDTEM
KEEQKSNTAGSQSQVETEA

# Quality evaluation based on protein interaction

**interacting proteins often have roughly similar abundances:**

origin recognition complex

| | |
|---|---|
| ORC1: 8.6 ppm | ORC4: 12.3 ppm |
| ORC2: 1.4 ppm | ORC5: 2.7 ppm |
| ORC3: 3.2 ppm | ORC6: 6.4 ppm |

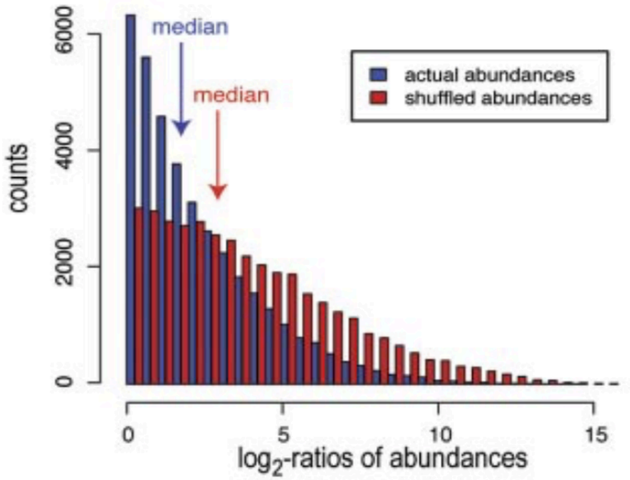
5.7 ± 4.2 ppm

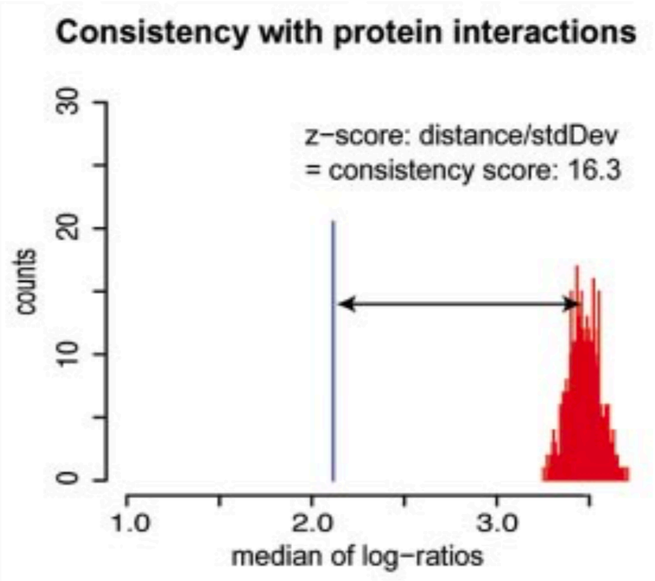
replication factor A

| |
|---|
| RFA1: 57 ppm |
| RFA2: 97 ppm |
| RFA3: 123 ppm |

92 ± 33 ppm

pairwise comparisons of all interacting proteins in yeast

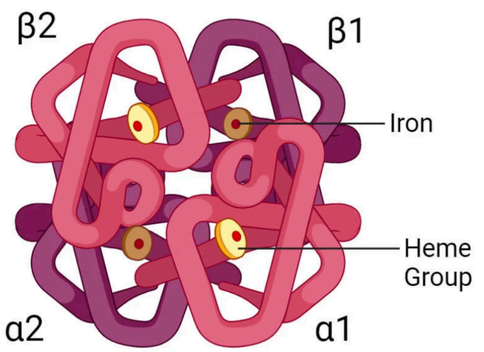


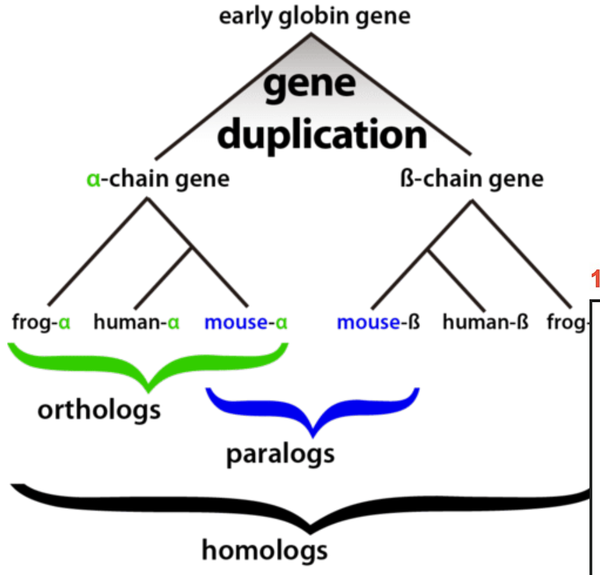



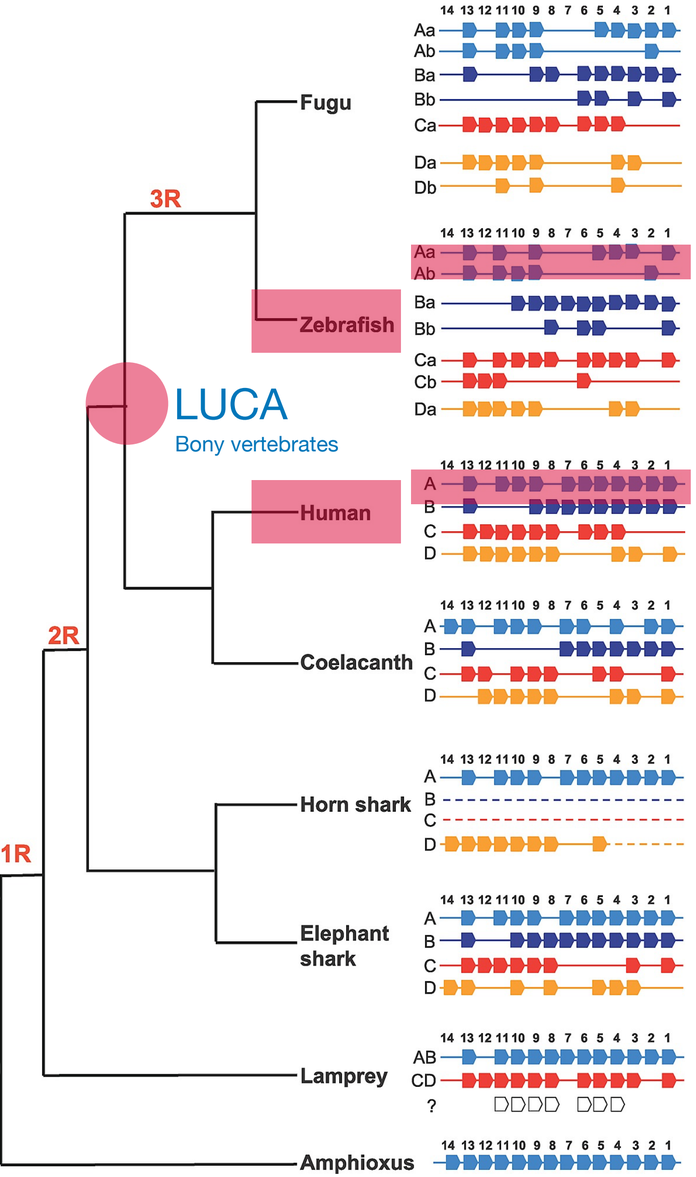

Wang et al, MCP 2012

# Ortholog relation

- Genes in different species that evolved from a common ancestral gene by speciation
- typically retain the same function
- mapped up to the last universal common ancestor (LUCA)



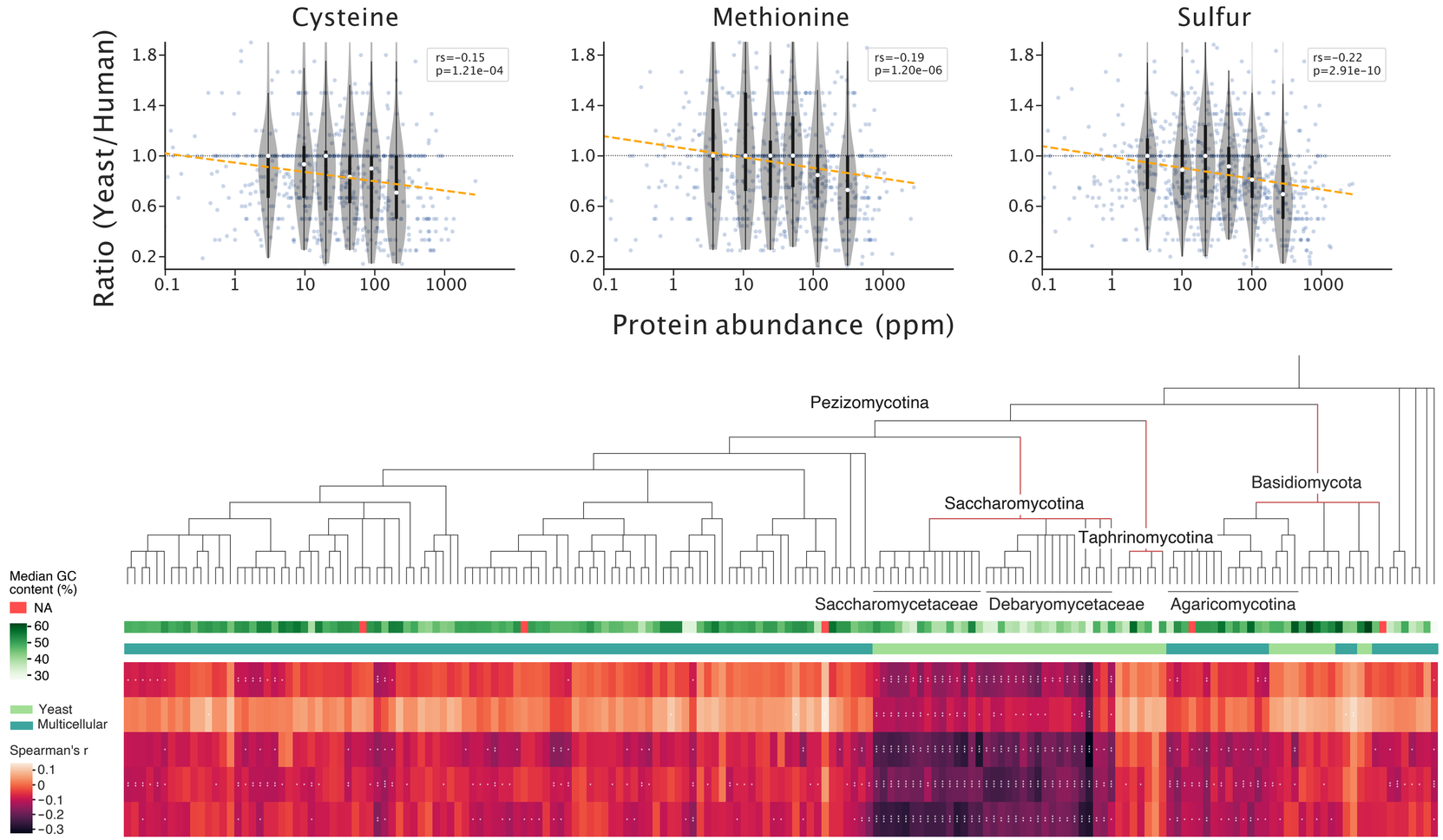




https://microbenotes.com/hemoglobin/

https://commons.wikimedia.org/wiki/File:Homology.png

Venkatesh et al., PLOS Biology 2015

# Protein abundance data reveals evolutionary signal



# Metadata

- Ontology
  - Species name → taxonomical ID
    - Human → 9606
  - Taxonomical level
    - Primates → 9443
    - Bacteria → 2
  - Tissue / organ → Uberon, Plant ontology …
    - THYROID_GLAND → UBERON:0002046
    - PERICARP → PO:0009084

# Metadata

- Ortholog mapping → COG IDs
  - APOA2 → 9443.ENOG504MJTR (primate level)
  - APOA2 → 33208.ENOG503BTNZ (metazoa level)
- Protein name → Protein ID
  - APOA2 → ENSP00000356969
- Publication
  - PMID, DOI, Journal, Year

# Access the data

## Bulk download

- https://pax-db.org/downloads/

### Index of /downloads/

../
1.0/
2.0/
2.1/
3.0/
4.0/
4.1/
4.2/
5.0/
6.0/
latest/

### Index of /downloads/latest/

| | | |
|---|---|---|
| ../ | | |
| datasets/ | 02-Oct-2025 08:39 | - |
| paxdb-mapped_peptides-v6.0/ | 05-Aug-2025 11:43 | - |
| paxdb-orthologs-v6.0/ | 04-Aug-2025 21:57 | - |
| paxdb-protein-sequences-v6.0/ | 03-Aug-2025 19:02 | - |
| paxdb-sdrf-v6.0/ | 18-Sep-2025 14:03 | - |
| paxdb-uniprot-links-v6.0/ | 05-Aug-2025 12:11 | - |
| paxdb-mapped_peptides-v6.0.zip | 06-Aug-2025 14:36 | 350439214 |
| paxdb-orthologs-v6.0.zip | 06-Aug-2025 14:37 | 39147658 |
| paxdb-protein-sequences-v6.0.zip | 06-Aug-2025 14:39 | 704363529 |
| paxdb-sdrf-v6.0.zip | 06-Aug-2025 14:39 | 94680 |
| paxdb-uniprot-crossreferences.txt | 05-Aug-2025 11:37 | 81955396 |
| paxdb-uniprot-links-v6.0.zip | 06-Aug-2025 14:37 | 16603448 |
| species_map_v5_v6.yml | 19-Jul-2025 07:52 | 2269 |
| unmapped_datasets_v5_v6.txt | 13-Aug-2025 16:15 | 400 |

# Access the data

## Filter and download individual dataset

Datasets

heart

| Name | Tissue type | Interaction consistency score | Coverage | Download |
|---|---|---|---|---|
| H.sapiens - Heart (Integrated) | Heart | 33.6 | 68% | Download |
| H.sapiens - Heart, Fetal, SC (Kim,nature,2014) | Heart | 26.7 | 51% | Download |
| H.sapiens - Heart, SC (Wangetal,molsystbiol2019) | Heart | 17.1 | 47% | Download |
| H.sapiens - Heart, SC (Peptideatlas,aug,2014) | Heart | 24.2 | 40% | Download |
| H.sapiens - Heart, SC (Kim,nature,2014) | Heart | 30.5 | 33% | Download |
| H.sapiens - Heart, SC (Aye,mol_bio_syst,2010) | Heart | 14.2 | 17% | Download |
| H.sapiens - Heart, SC (Kline,j.proteome_res,2008) | Heart | 9.3 | 17% | Download |
| H.sapiens - Heart, normalized data APEX (Aye,mol_bio_syst,2010) | Heart | 13 | 11% | Download |

# Access the data

## API

1. Show info of all datasets of *Arabidopsis thaliana*.
2. What are all the information about dataset xxx?
3. What are all protein abundance and annotation in the dataset xxx?
4. How is the protein abundance distrubution of dataset xxx?
5. Where does the protein xxx stand in the distribution?
6. What are all abundances of the protein by string ID xxx?
7. What are all abundances of the protein by Ubiprot ID xxx?
8. What are abundances of multiple proteins x, y, z … across all datasets?

1. https://api.pax-db.org/species/3702
2. https://api.pax-db.org/dataset/9606/986013392/abundances
3. https://api.pax-db.org/dataset/9606/986013392/
4. https://api.pax-db.org/dataset/986013392/histogram/
5. https://api.pax-db.org/dataset/986013392/histogram/?highlightProteinId=ENSP000003700
6. https://api.pax-db.org/protein/string/9606.ENSP00000295897
7. https://api.pax-db.org/protein/uniprot/Q851P9_ORYSJ
8. https://api.pax-db.org/proteins?ids=9606.ENSP00000269305,9606.ENSP00000258149

# Access the data

## API

1. Show info of all datasets of *Arabidopsis thaliana*.

1. https://api.pax-db.org/species/3702


```
{
    "id": 3702,
    "name": "Arabidopsis thaliana",
    "compact_name": "Arabidopsis thaliana",
    "num_proteins": 27413,
    "genome_source": "<a href='http://plants.ensembl.org/Arabidopsis_thaliana/Info/
    "genome_source_version": " release 56",
    "datasets": [
        {
            "id": 1318907283,
            "name": "A.thaliana - Leaf, Short day optimal water, SC (Baerenfaller,m
            "score": 6.7,
            "description": "abundance based on Spectral counting SOW, Short Day Opt
            "organ": "LEAF",
            "integrated": false,
            "coverage": 9,
            "publication_year": 2012,
            "num_abundances": 2438,
            "hasPeptideCounts": true,
            "filename": "3702-2.1ArabidopsisLeaf_waterDefict_SOW.txt"
```

# Engaging the community

- Submit issues - data error
- upload their own data - contribute to DB
- Submit processing request providing metadata
  - off-load curation to community / domain expert
  - Github Issues Tracker (visible)

# Summary

- Why FAIR?
  - Data scattered across publications and formats can reveal much more when analyzed systematically
- How?
  - Persistent identifier, rich metadata (findable)
  - Versioned data archive, cater to different use cases (accessible)
  - Metadata standardization (Interoperable)
  - Data provenance, license (re-usable)