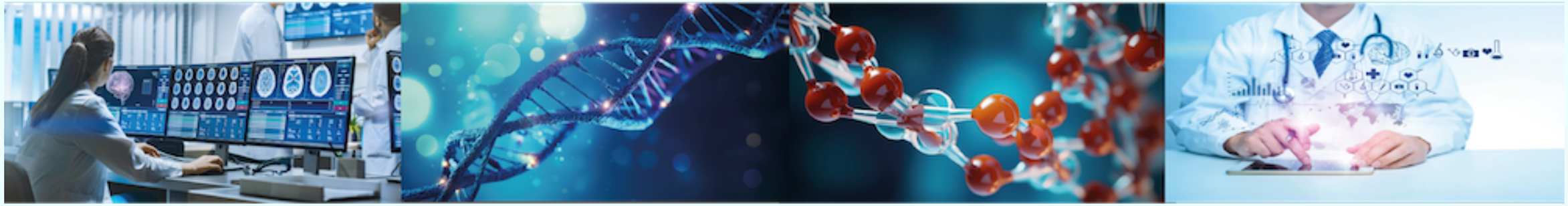


Programming for Biomedical Informatics

Lecture 3 - Introduction to the Biomedical Dataverse

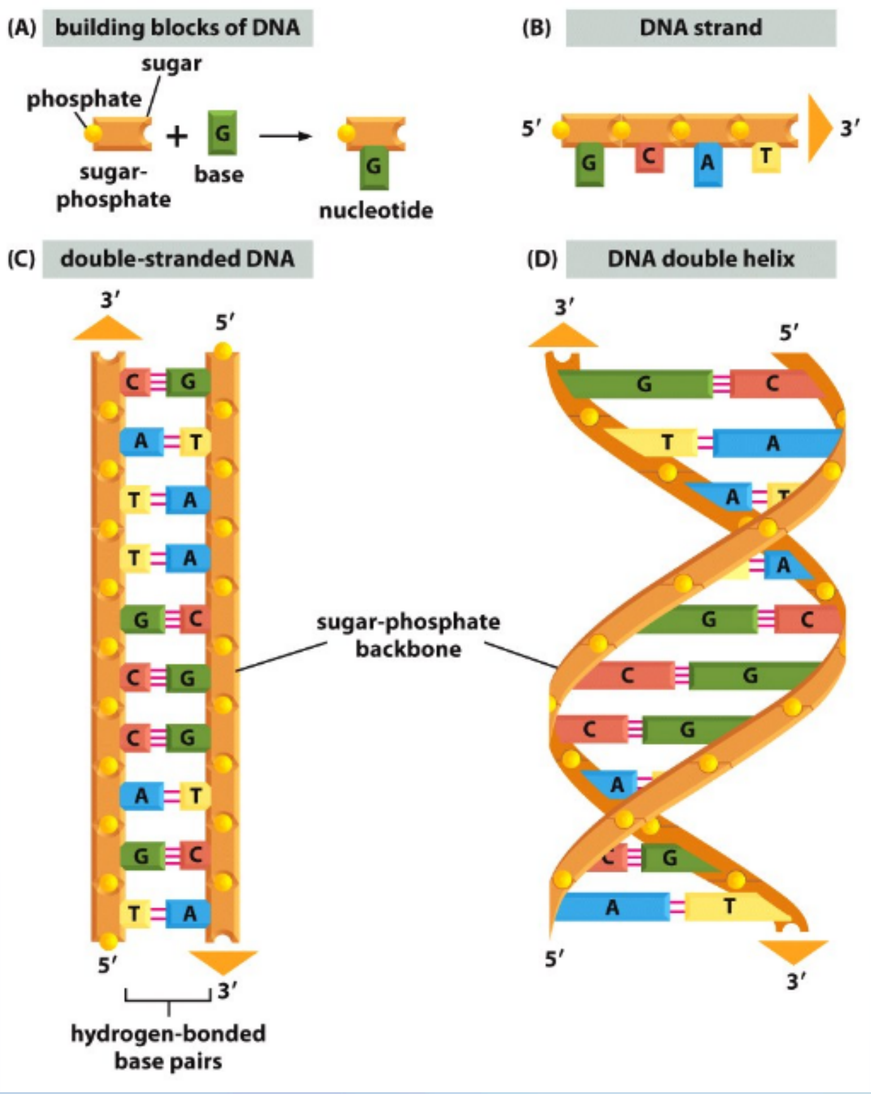
https://github.com/biomedical-informatics/pbi

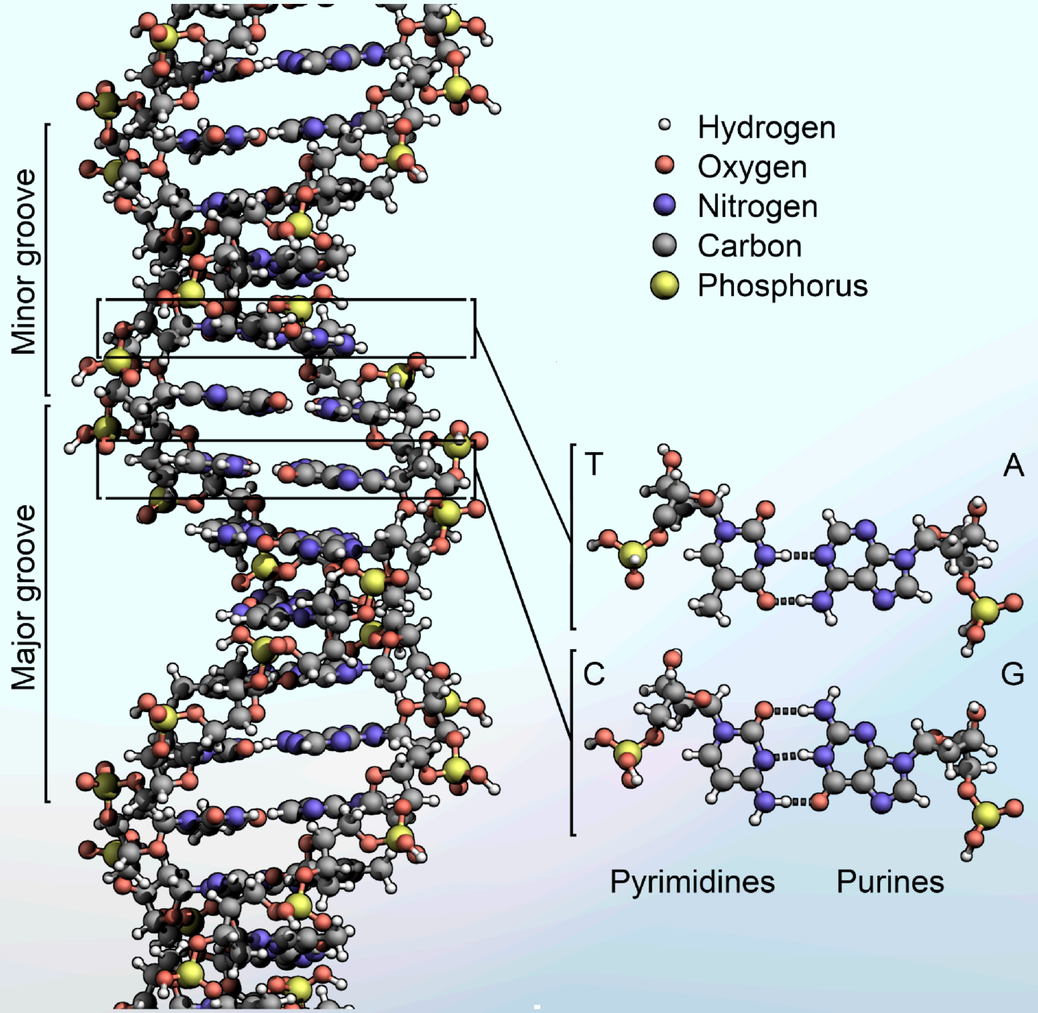
Ian Simpson
ian.simpson@ed.ac.uk

THE UNIVERSITY of EDINBURGH
informatics

Primer on Central Dogma (optional)

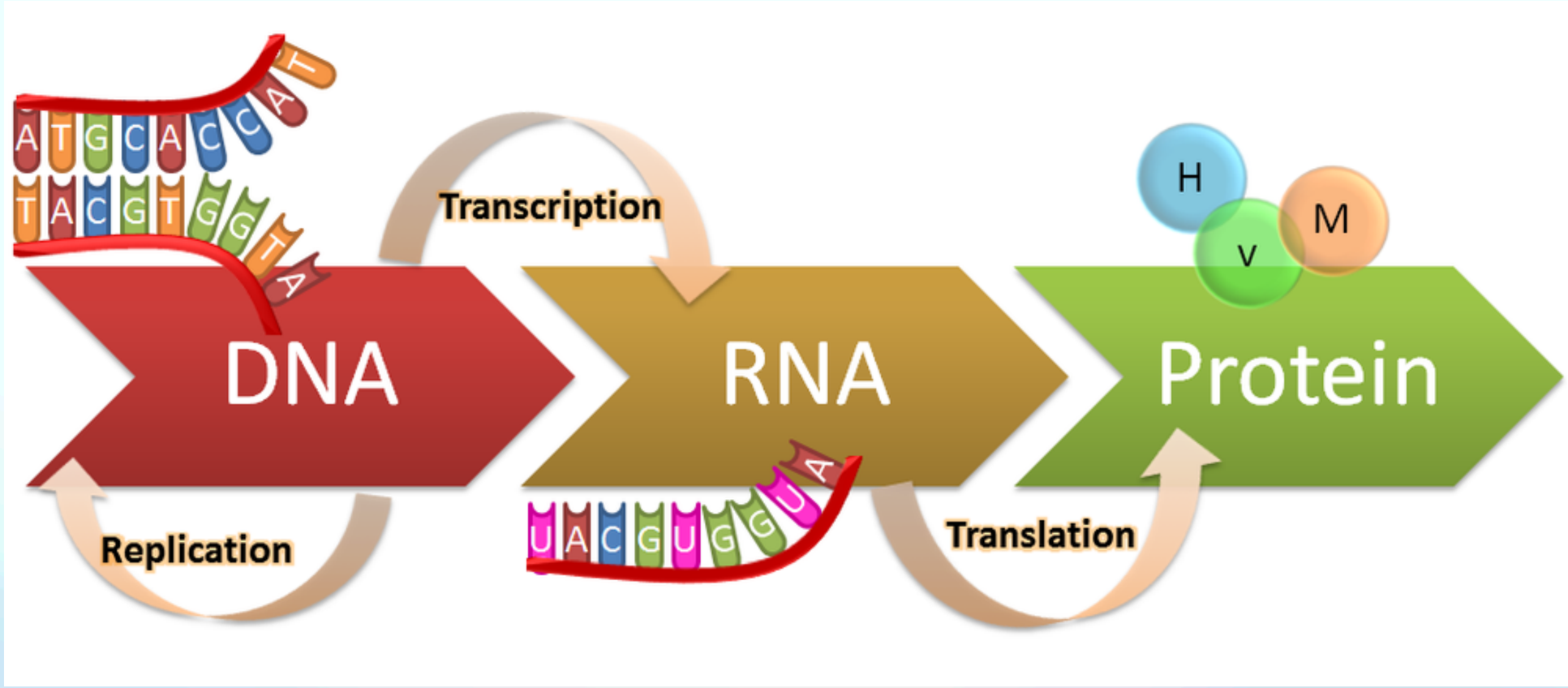
Deoxy-ribo**N**ucleic **A**cid (DNA)



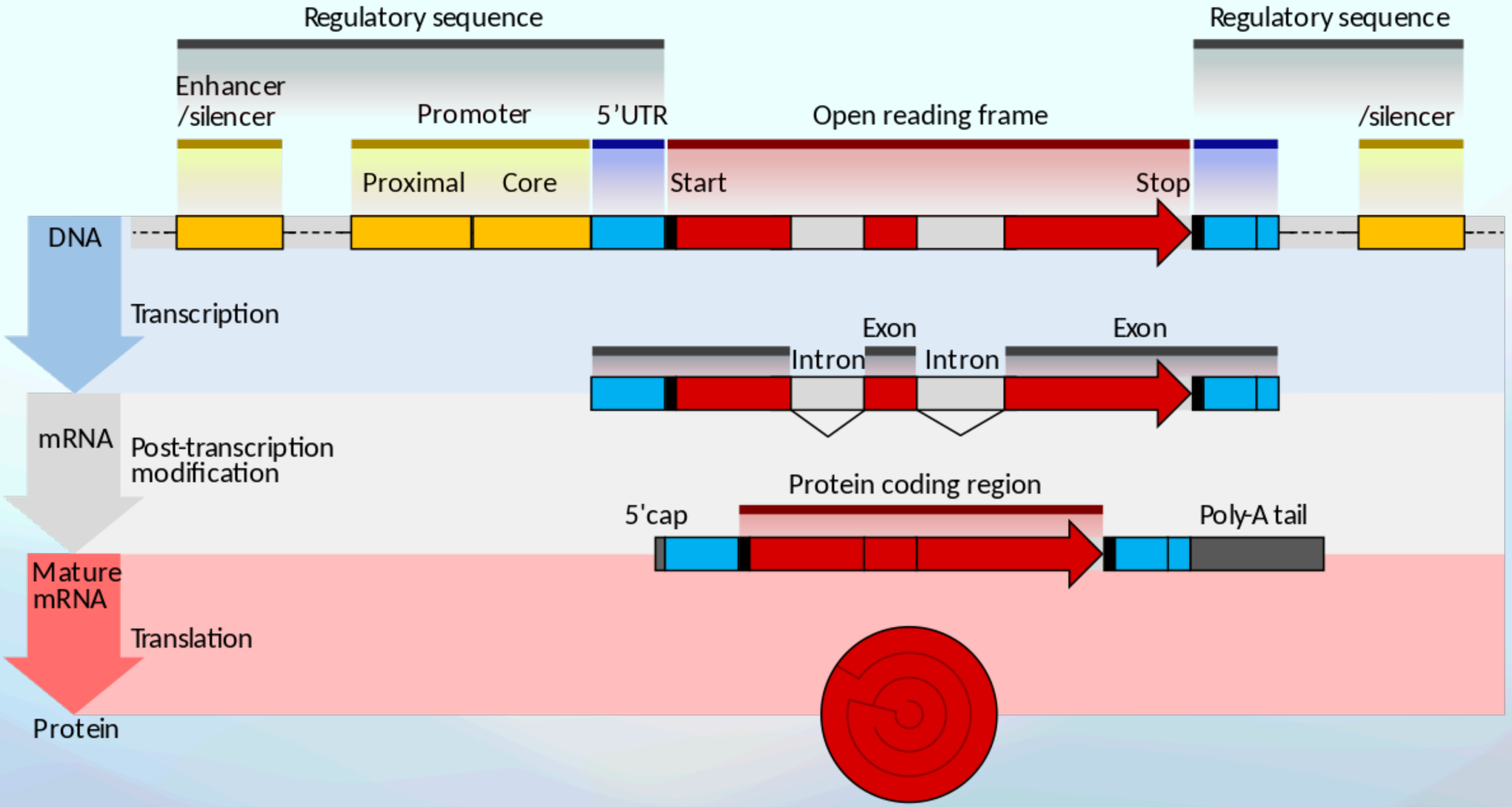


A - adenine
C - cytosine
G - guanine
T - thymine
U - uracil

The “Central Dogma”



The “Central Dogma”



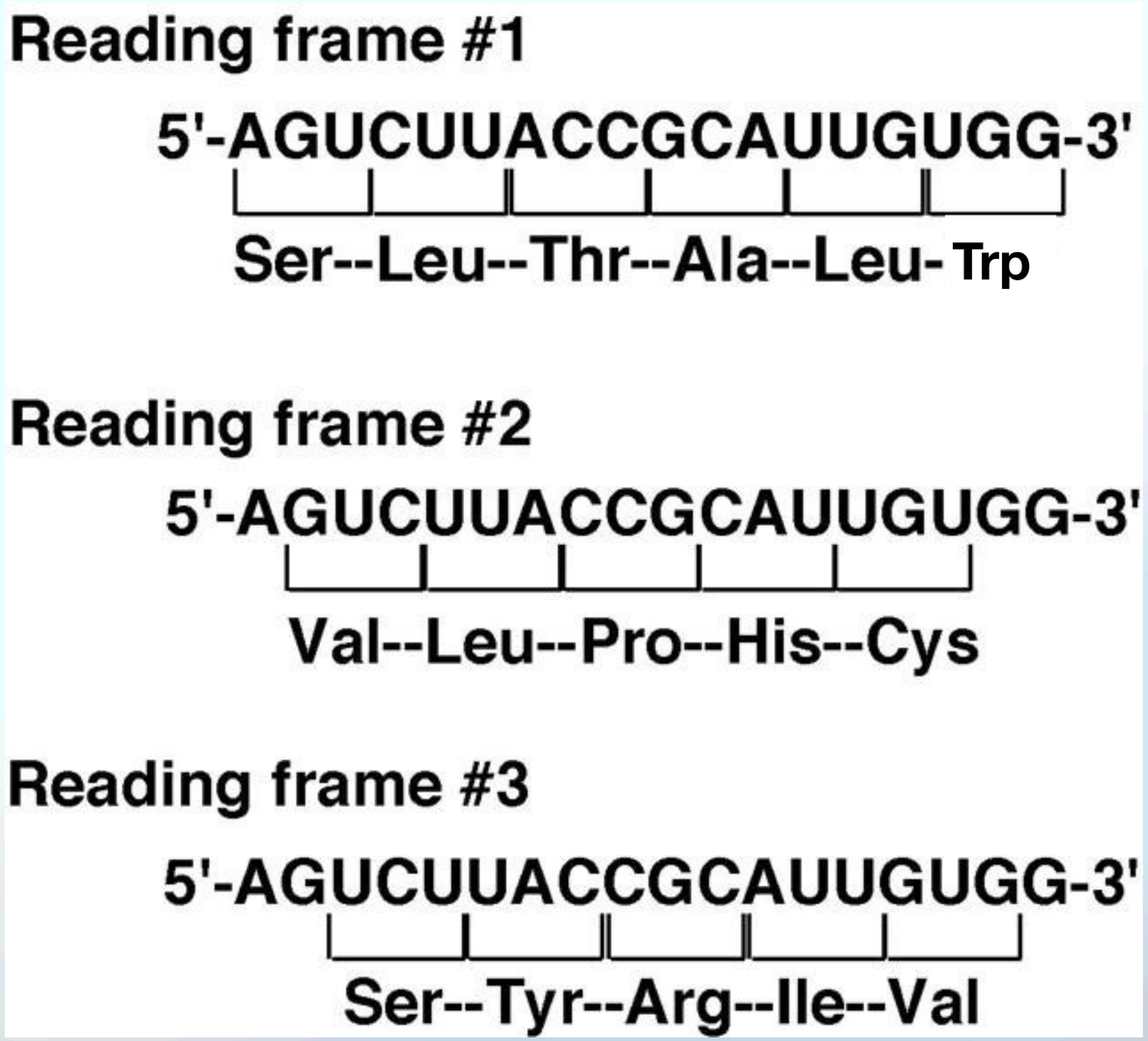
The Genetic Code

Standard genetic code

1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*)	TGA	Stop (*Opal*)	A
	TTG		TCG		TAG	Stop (*Amber*)	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

Reading mRNA Sequences

Start: AUG
Stop: UAA, UAG, UGA



Sources of Biomedical Data

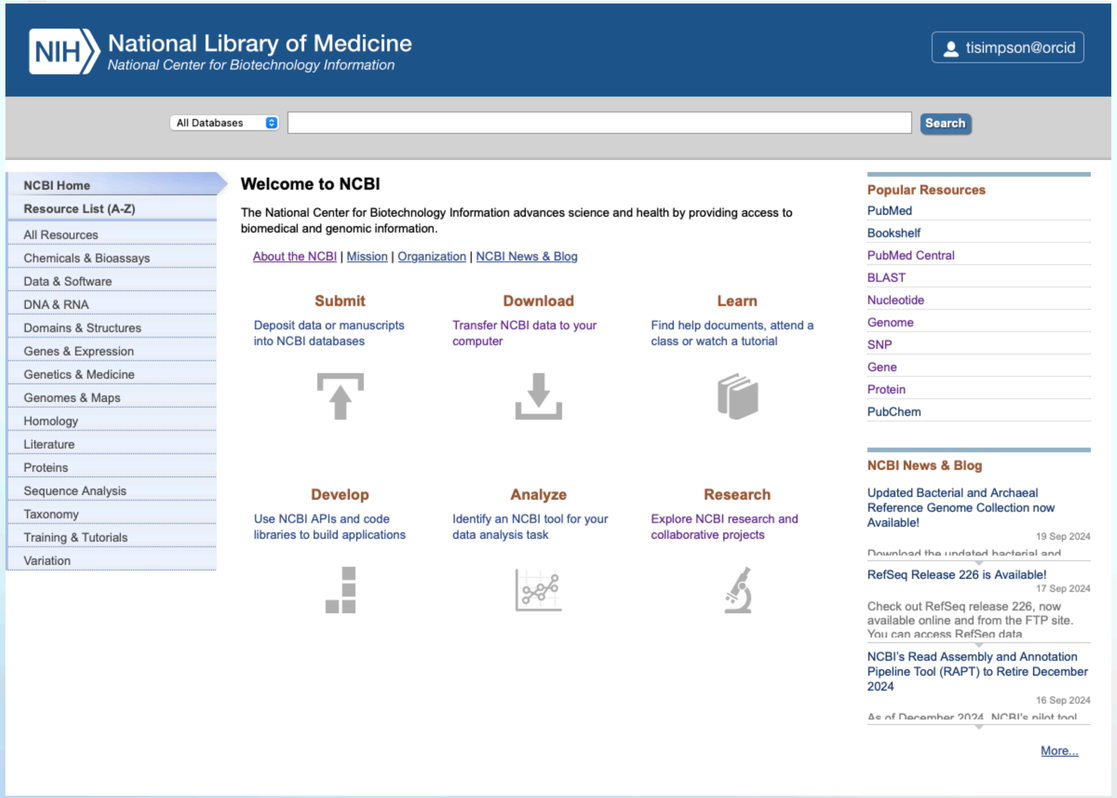
- There are many places where data relevant to biomedical research can be found
- The data can be roughly split into:
  - **sequences & structures** (DNA, RNA, proteins, metabolites, oligonucleotides, peptides, sugars, drugs…)
  - **images** (photographic, histological, CT, MRI, PET, ultrasound, X-ray, microscopy…)
  - **reference** (literature repositories, domain specific information databases, standards, data structures…)
  - **health** (clinical notes, reports, patient & cohort meta-data, prescription, surveys/questionnaires, assays…)
- The availability, coverage, and quality of these data vary hugely
- Much work you will do in this domain requires a detailed understanding of the primary data:
  - what do entities represent?
  - how are they related?
  - what information do they capture? (and as importantly what don't they capture)
  - is the data noisy, partial (at random or otherwise), or biased?
- If these (and other) issues are not addressed at the beginning, much that follows is flawed

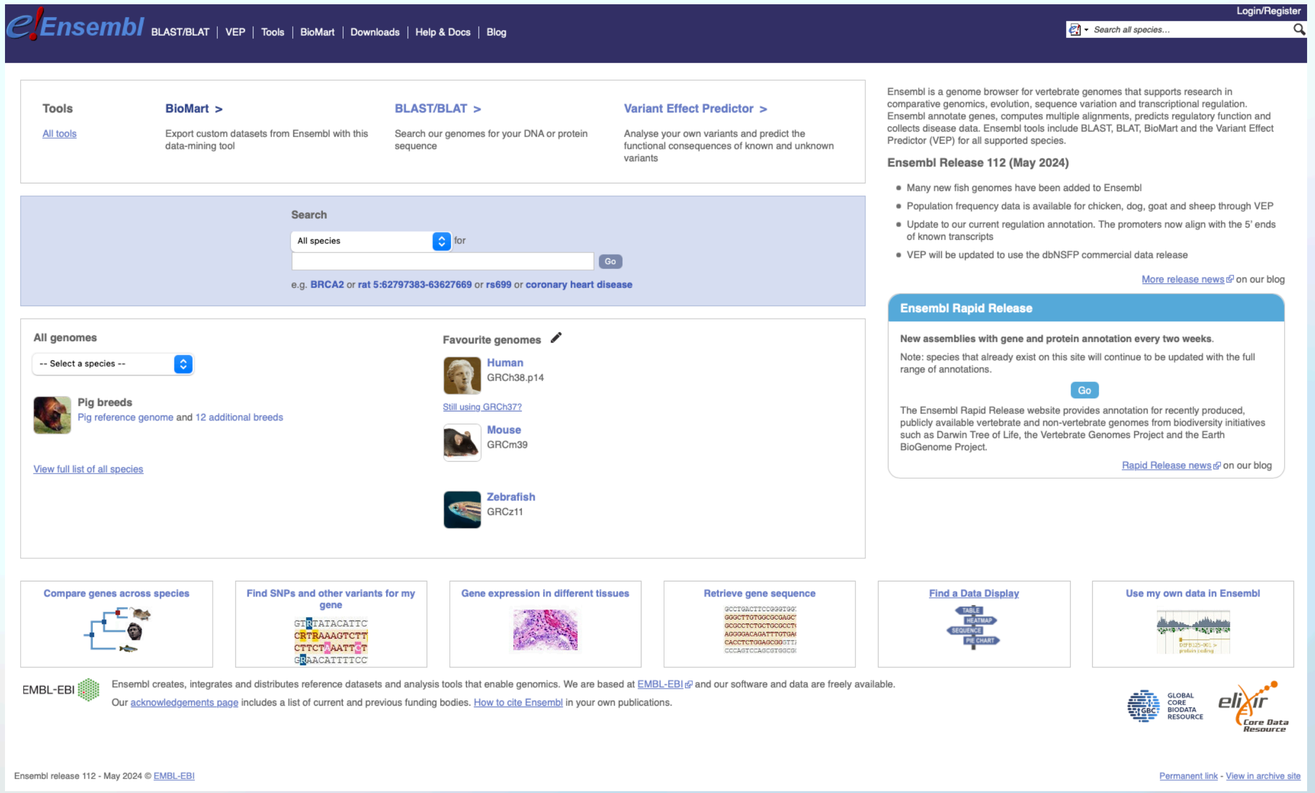
Sources of Biomedical Data

- In addition to the types of data they store data is commonly split into that which is openly available and that which is only available under restricted conditions.
- Data sources that provide “reference” information (for example a catalogue of all the genes in a human genome and the proteins they (may) produce) tend to be open source and comply with internationally agreed standards for how they are structured.
- Data sources that that contain personal and/or proprietary data are, unsurprisingly, limited access and may not adhere to external standards (though many do, and/or provide descriptors of how data and records are structured).
- Healthcare data is almost exclusively protected, often access is granted after training, accreditation, project approval, and ethical consent. Data is commonly stored in a “safe-haven” environment where individual level data can never leave. Analyses take place within the protected environment.
- There are a small number of openly available datasets that do contain individual level data in a suitably anonymised form. We will use some of these during the course.

Sources of Sequence & Structural Data

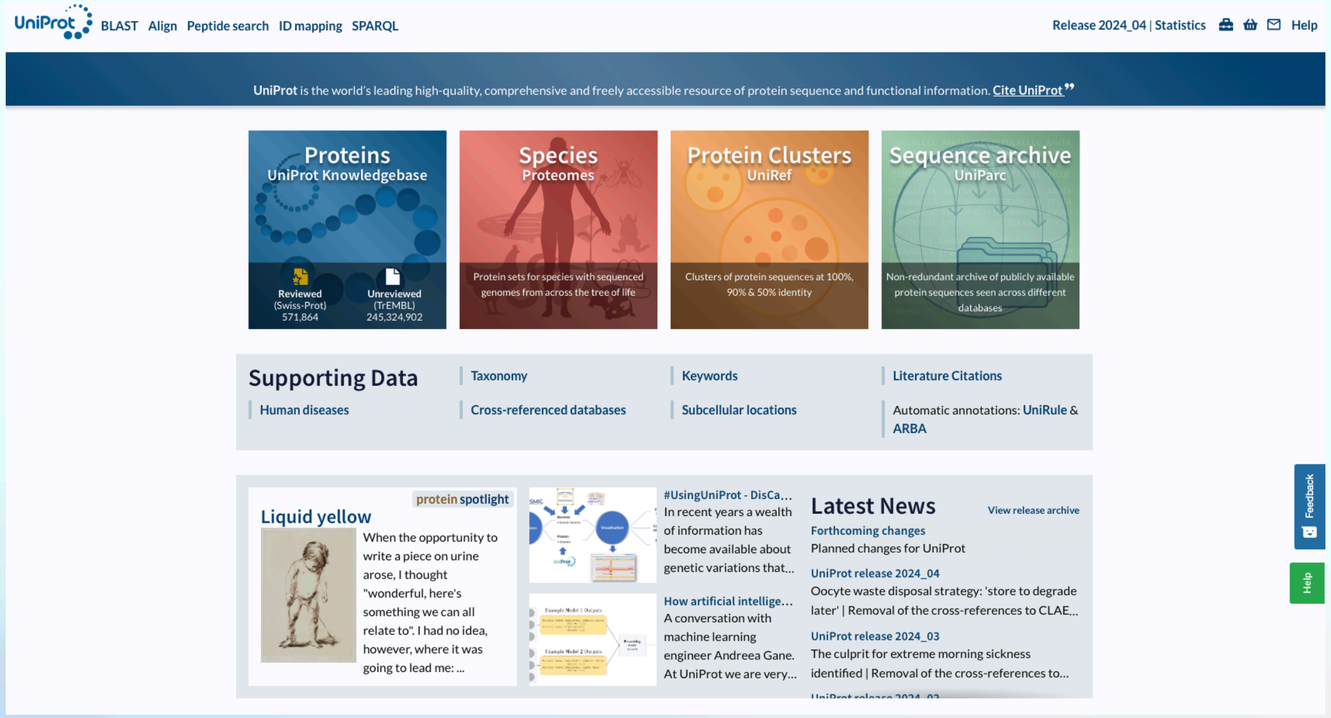
NCBI (National Centre for Biotechnology Information) - https://www.ncbi.nlm.nih.gov/
Ensembl (EMBL-EBI) - https://www.ensembl.org/

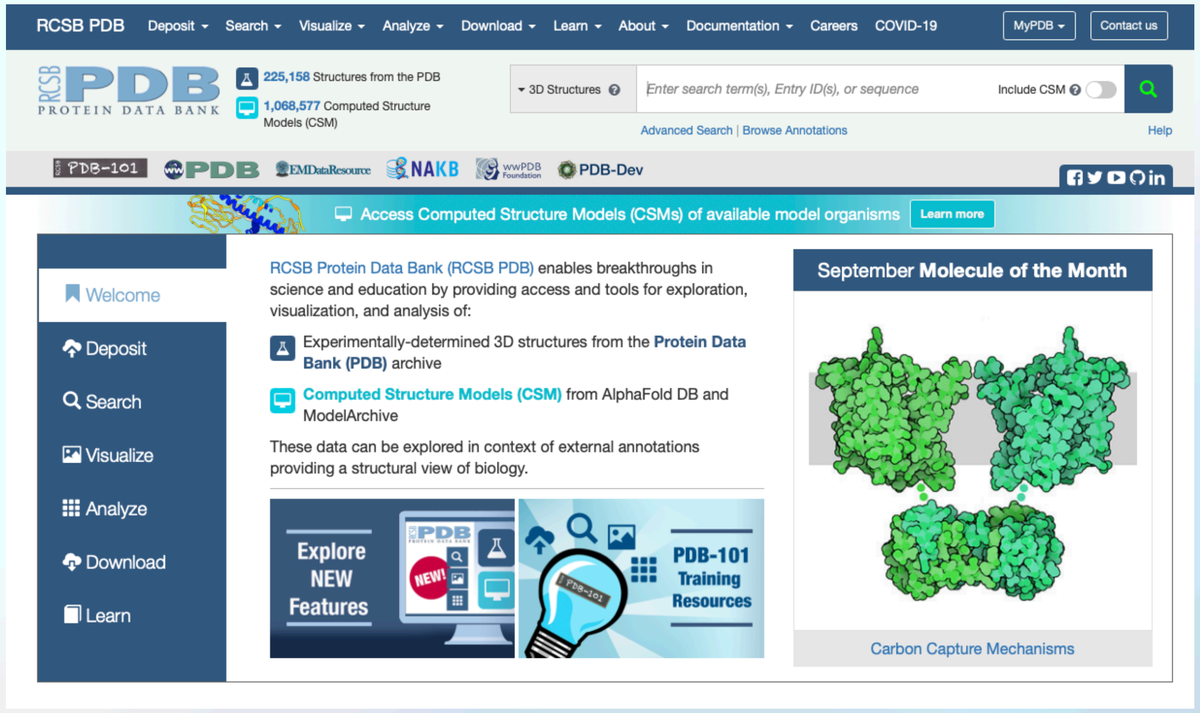




Sources of Sequence & Structural Data

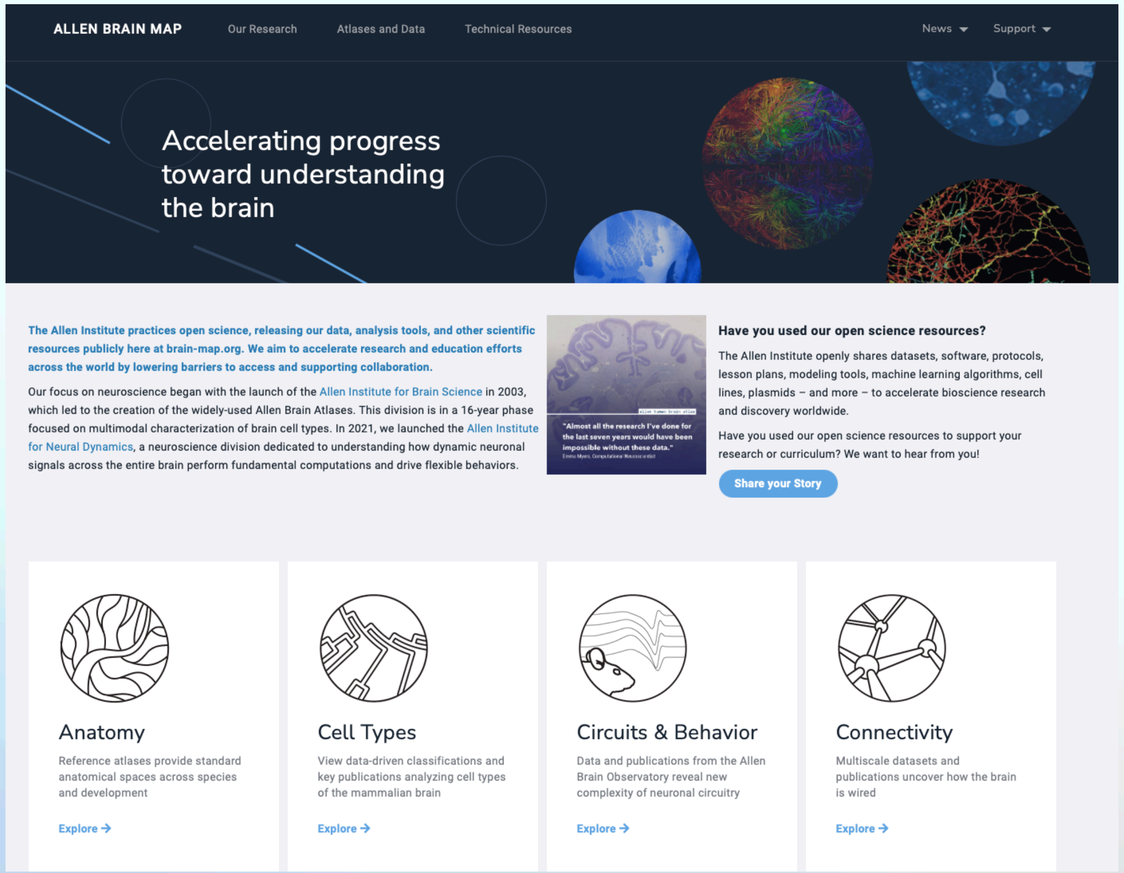
UniProt (Unified Proteins) - https://www.uniprot.org/
PDB (Protein Data Bank) - https://www.rcsb.org/





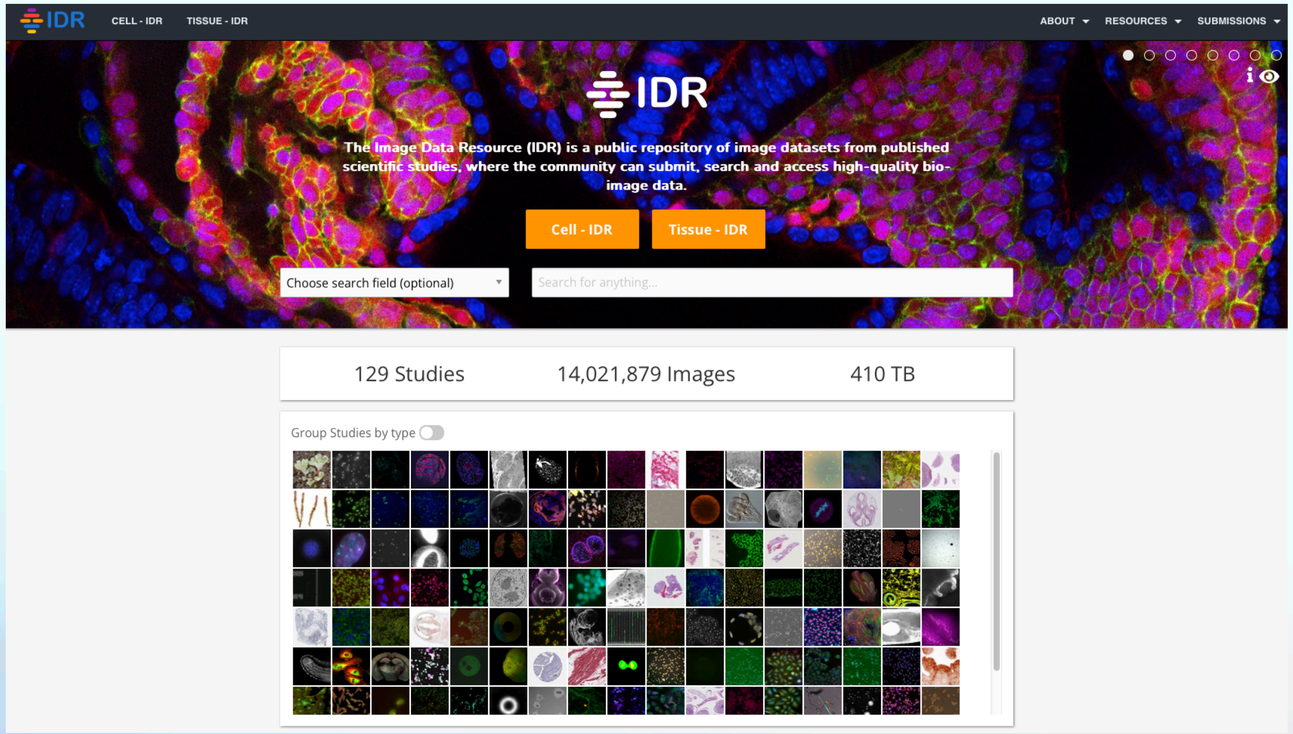
Sources of Imaging Data

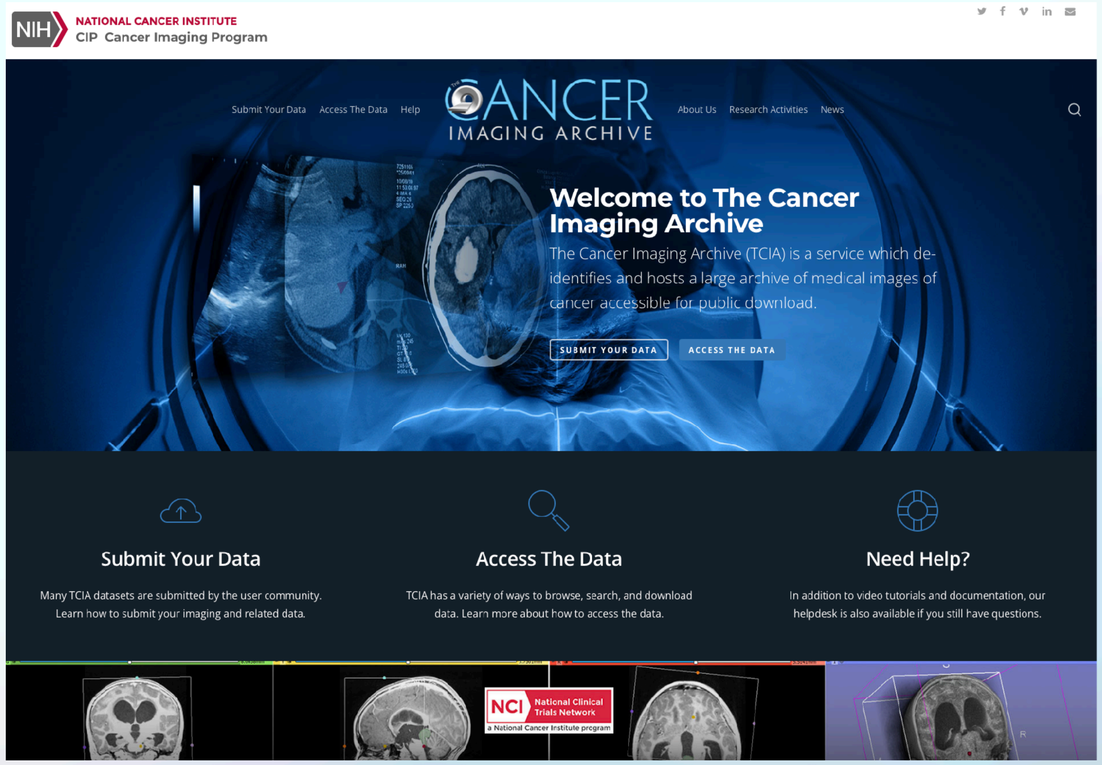
Allen Brain Atlas - https://www.brain-map.org/
Human Protein Atlas - https://www.proteinatlas.org/



Sources of Imaging Data

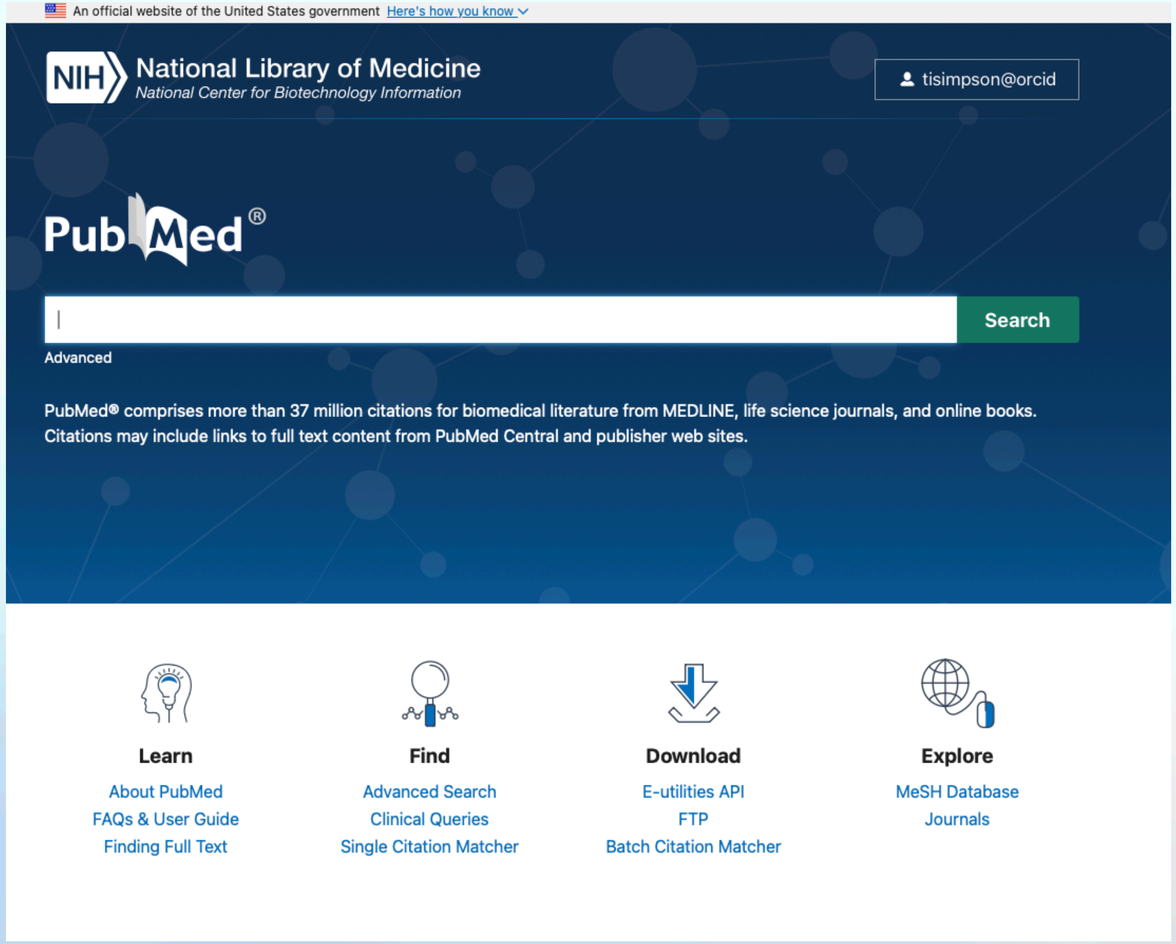
Image Data Resource - https://idr.openmicroscopy.org/
The Cancer Imaging Archive - https://www.cancerimagingarchive.net/

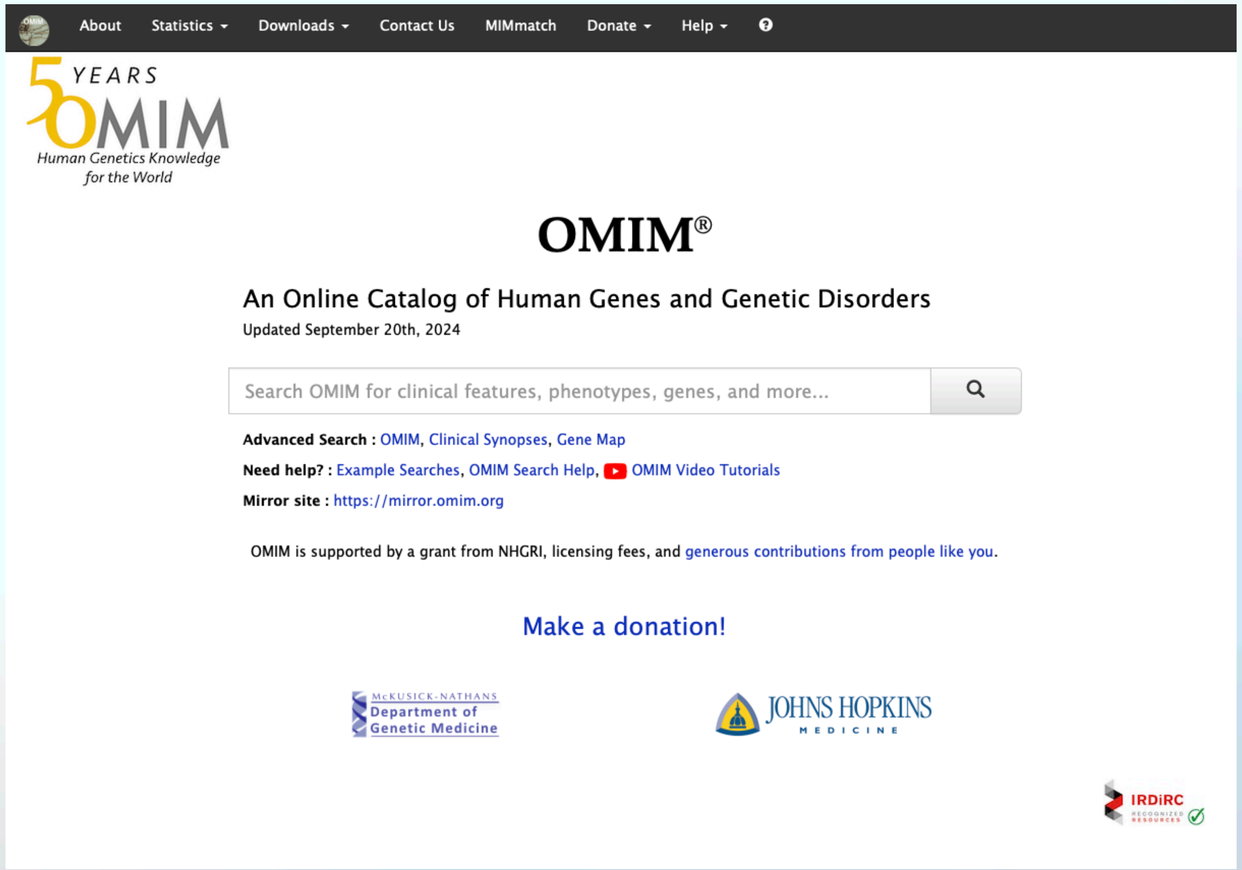




Sources of Biomedical Reference Data

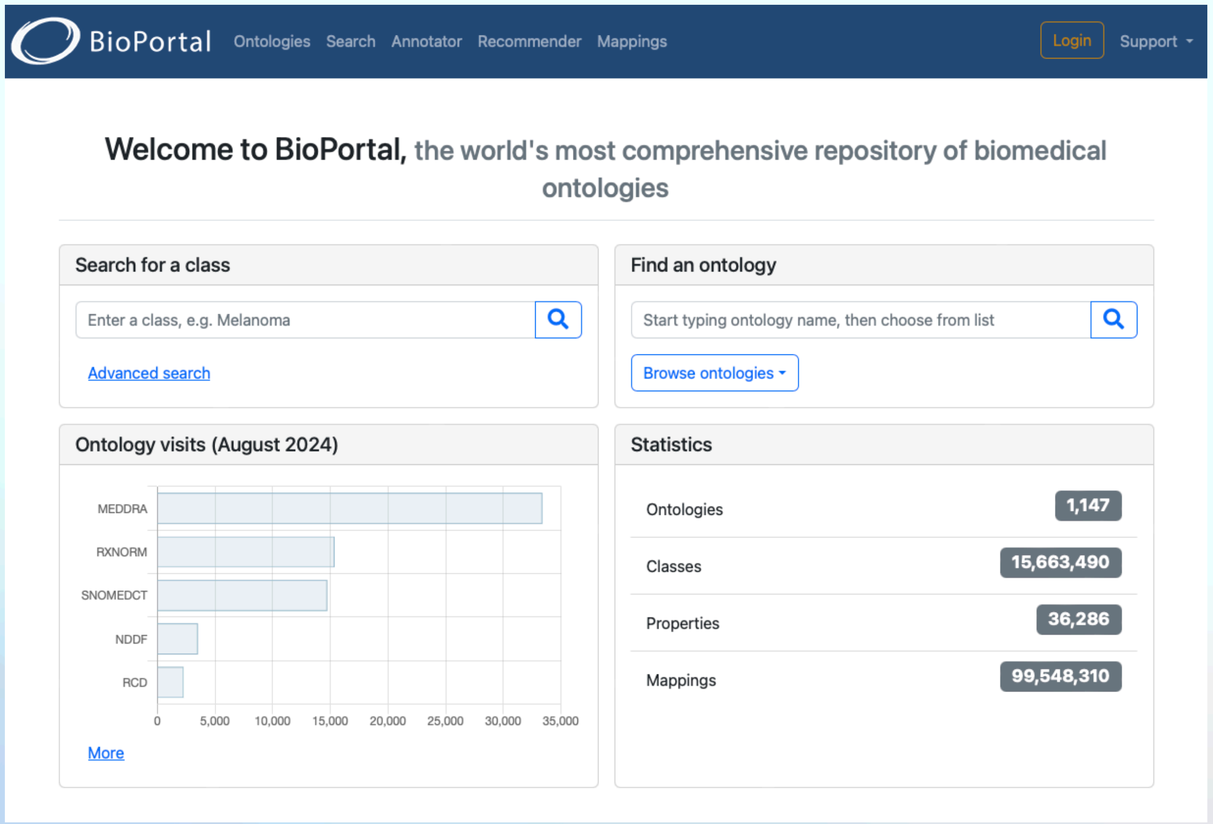
PubMed - https://pubmed.ncbi.nlm.nih.gov/
OMIM (Online Mendelian Inheritance in Man) - https://www.omim.org/

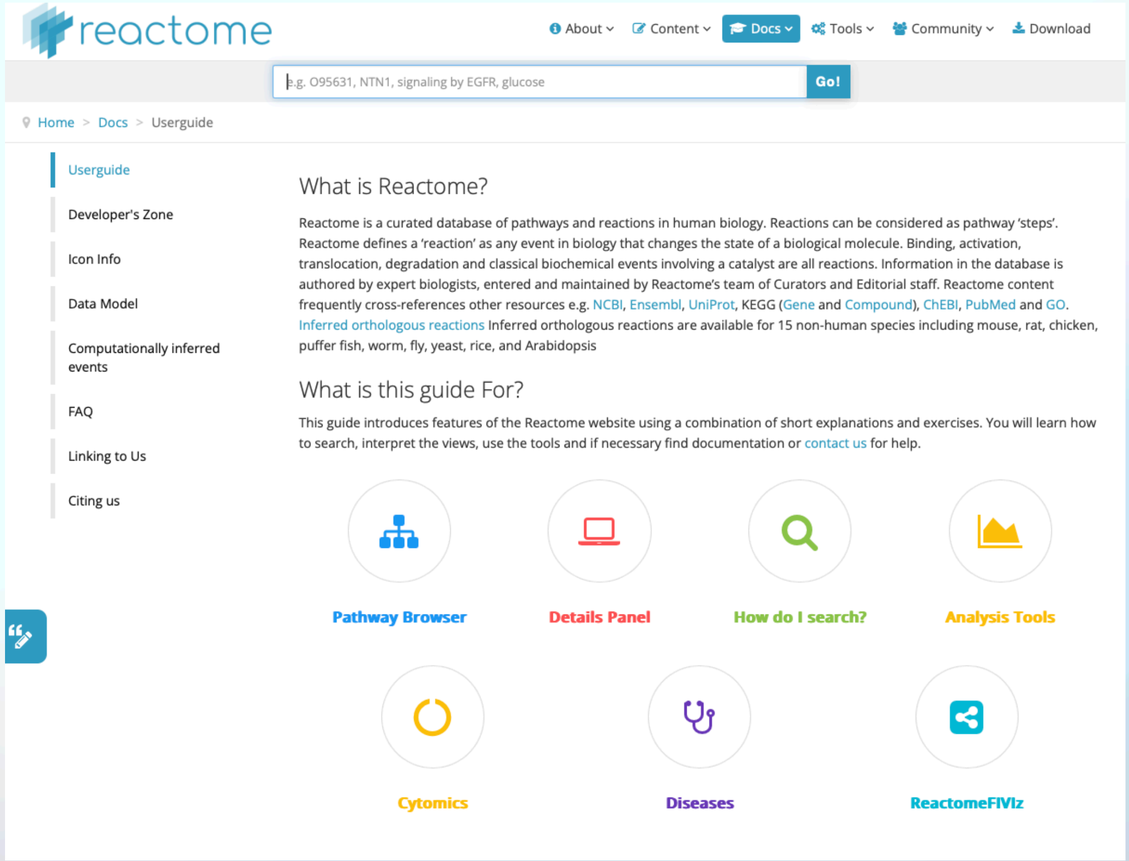




Sources of Biomedical Reference Data

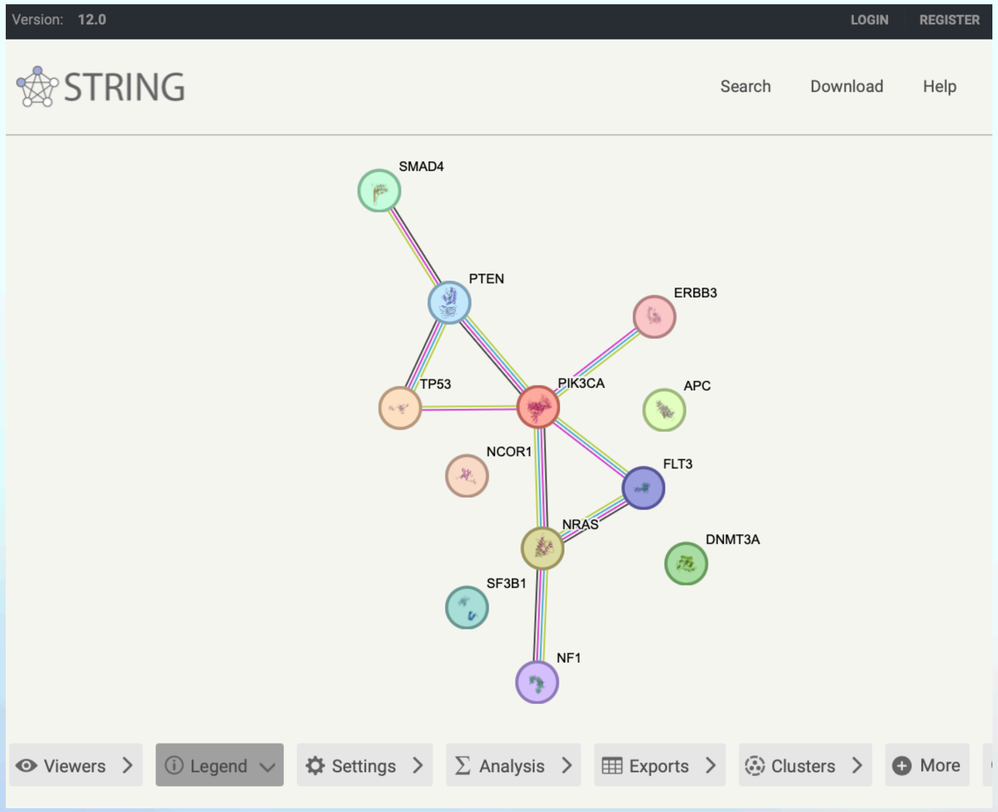
NCBO Bioportal -https://bioportal.bioontology.org/
Reactome - https://reactome.org/

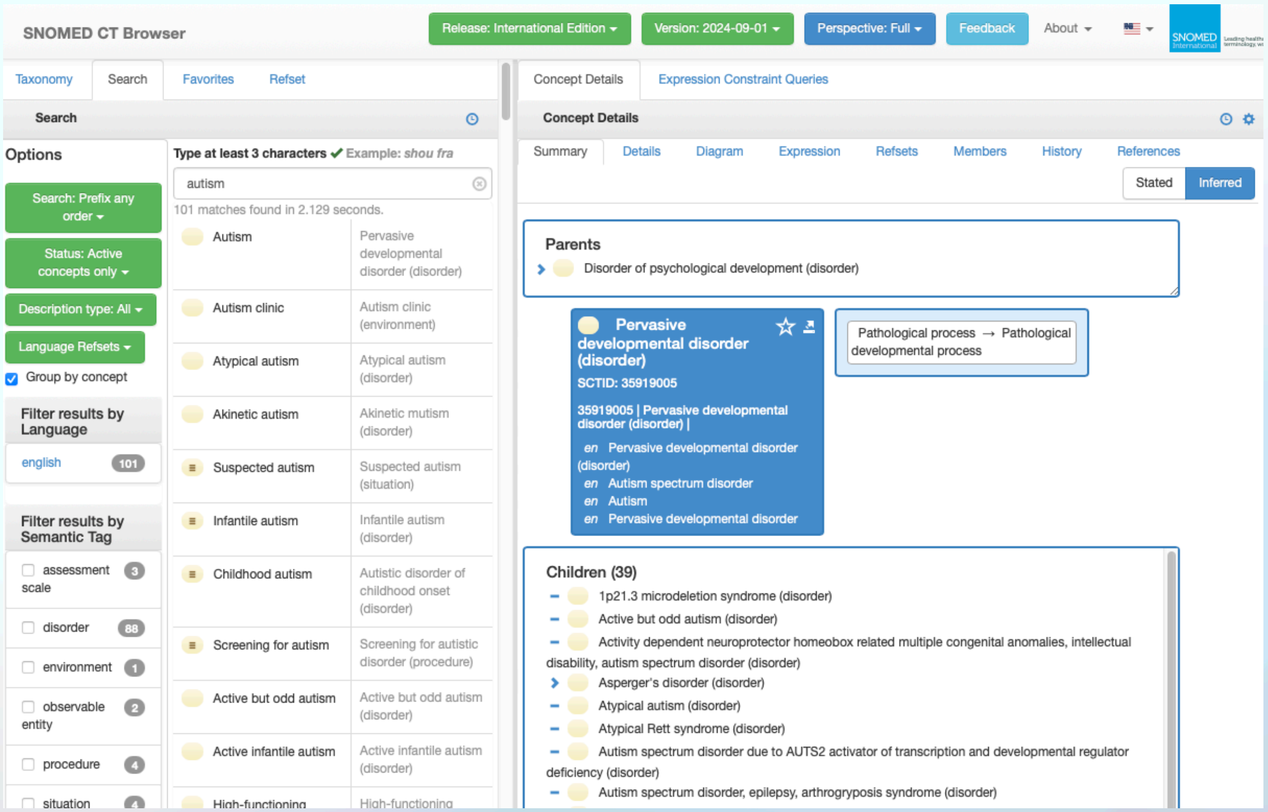




Sources of Biomedical Reference Data

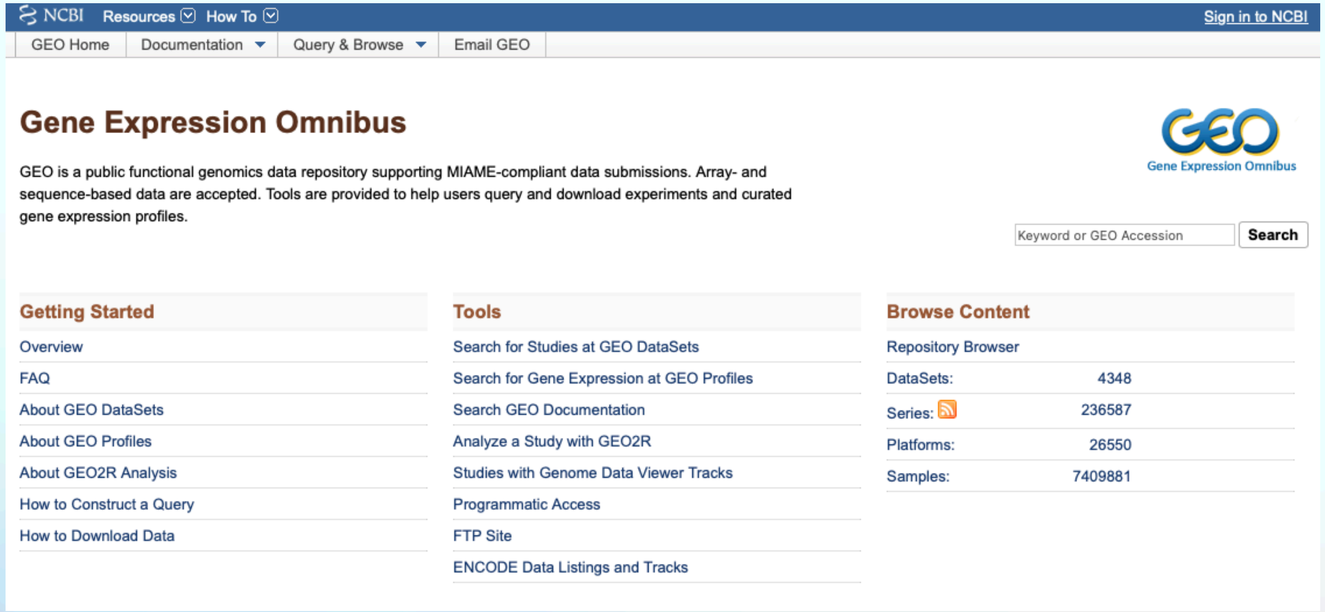
STRING - https://string-db.org (Protein Interaction/Association Database)
SNOMED-CT - https://uts.nlm.nih.gov/uts/ (UMLS - Unified Medical Language System)

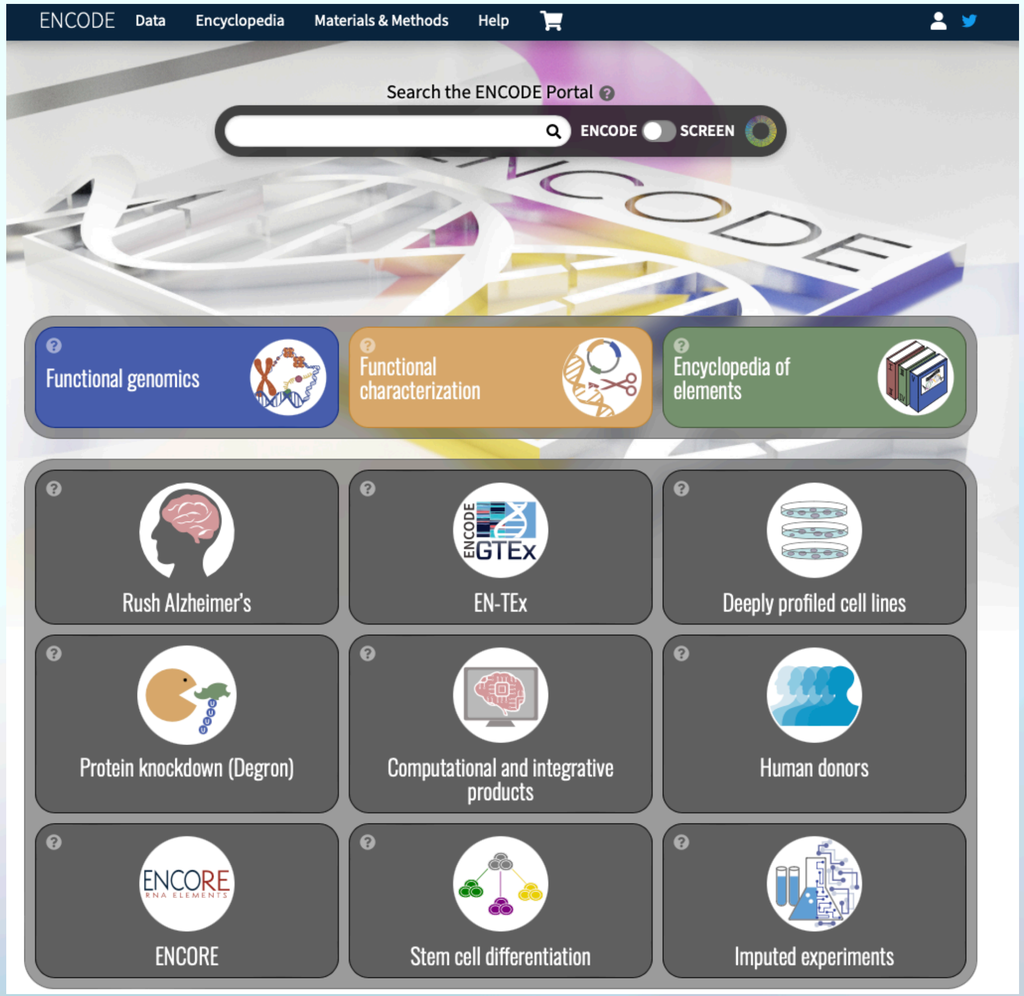




Sources of Biomedical Experimental Data

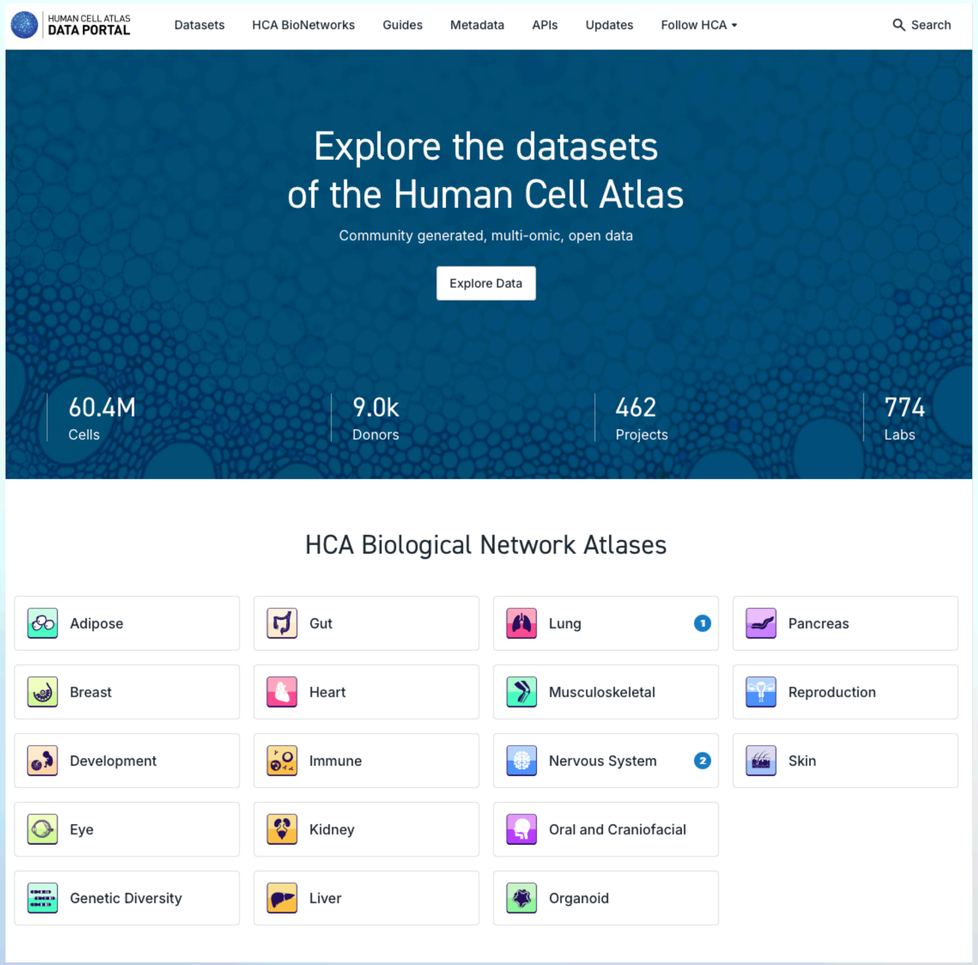
GEO (Gene Expression Omnibus) - https://www.ncbi.nlm.nih.gov/geo/
ENCODE (Encyclopaedia of DNA elements) - https://www.encodeproject.org/





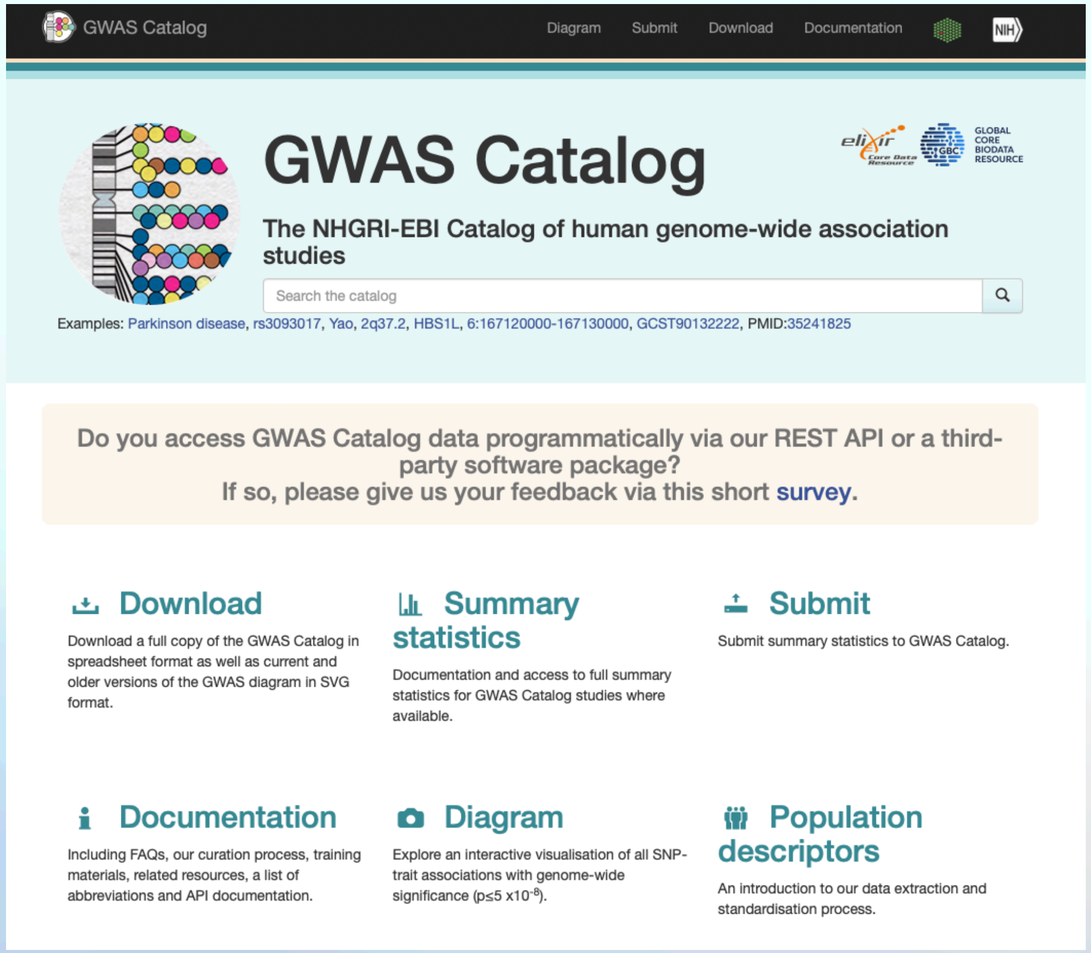
Sources of Biomedical Experimental Data

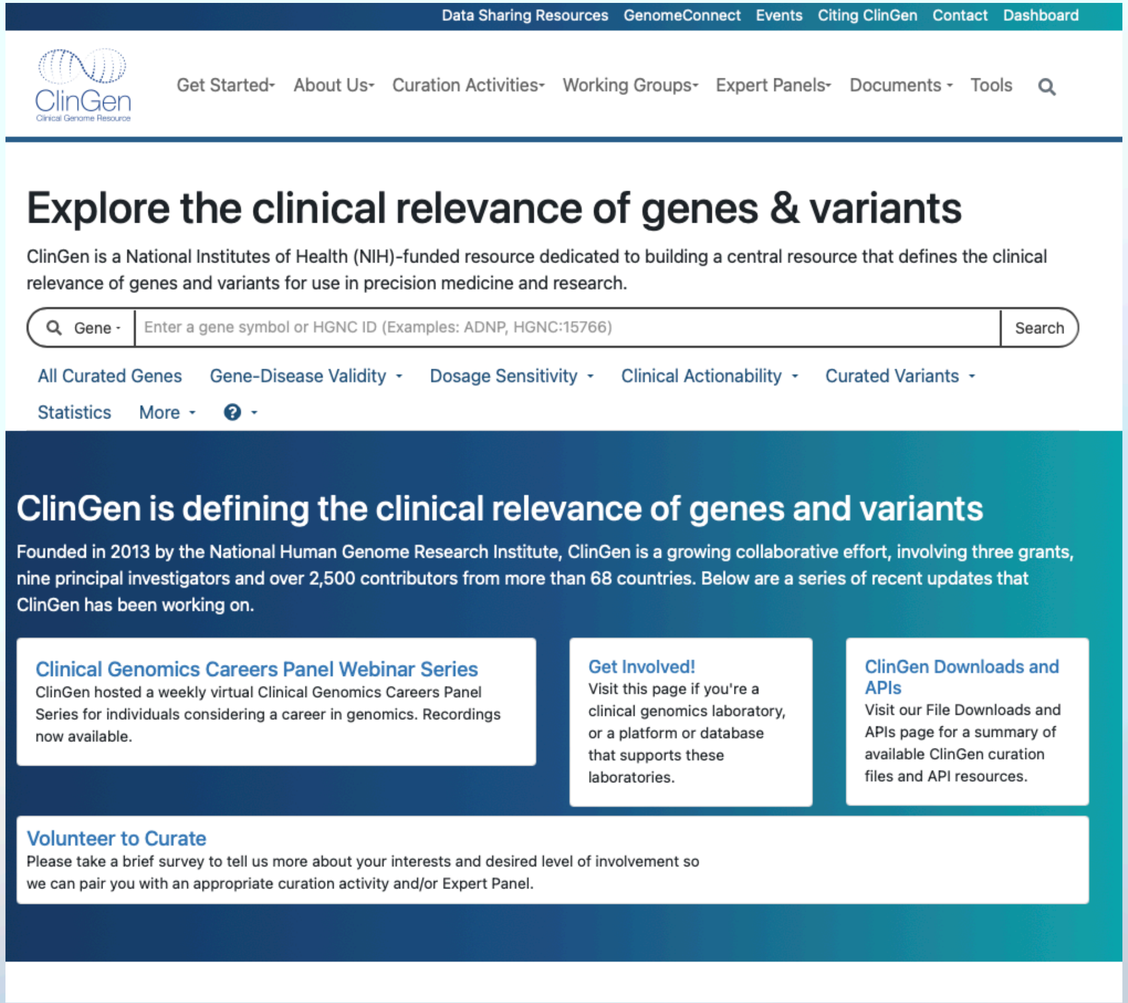
HCA (Human Cell Atlas) - https://www.humancellatlas.org
GTEx Portal (Adult Genotype Expression Project) - https://gtexportal.org/



Sources of Biomedical Experimental Data

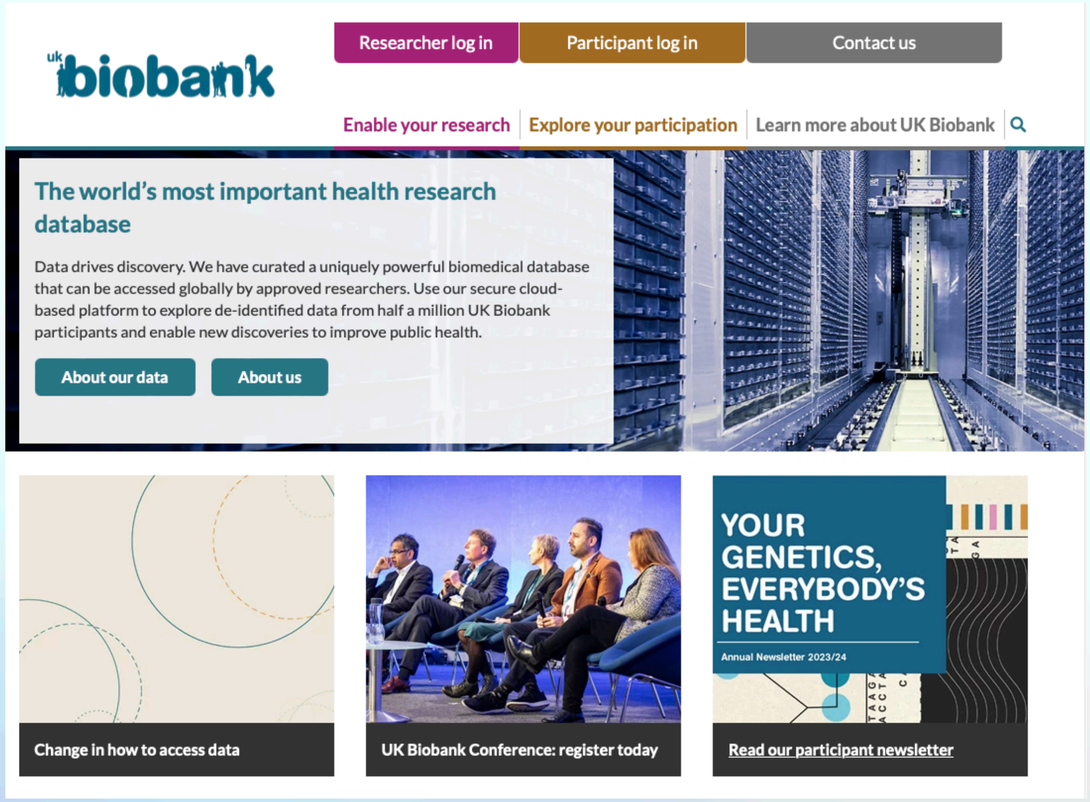
GWAS Catalog (Genome Wide Association Studies) - https://www.ebi.ac.uk/gwas/
ClinGen (The Clinical Genome - variation/disease) - https://www.clinicalgenome.org/

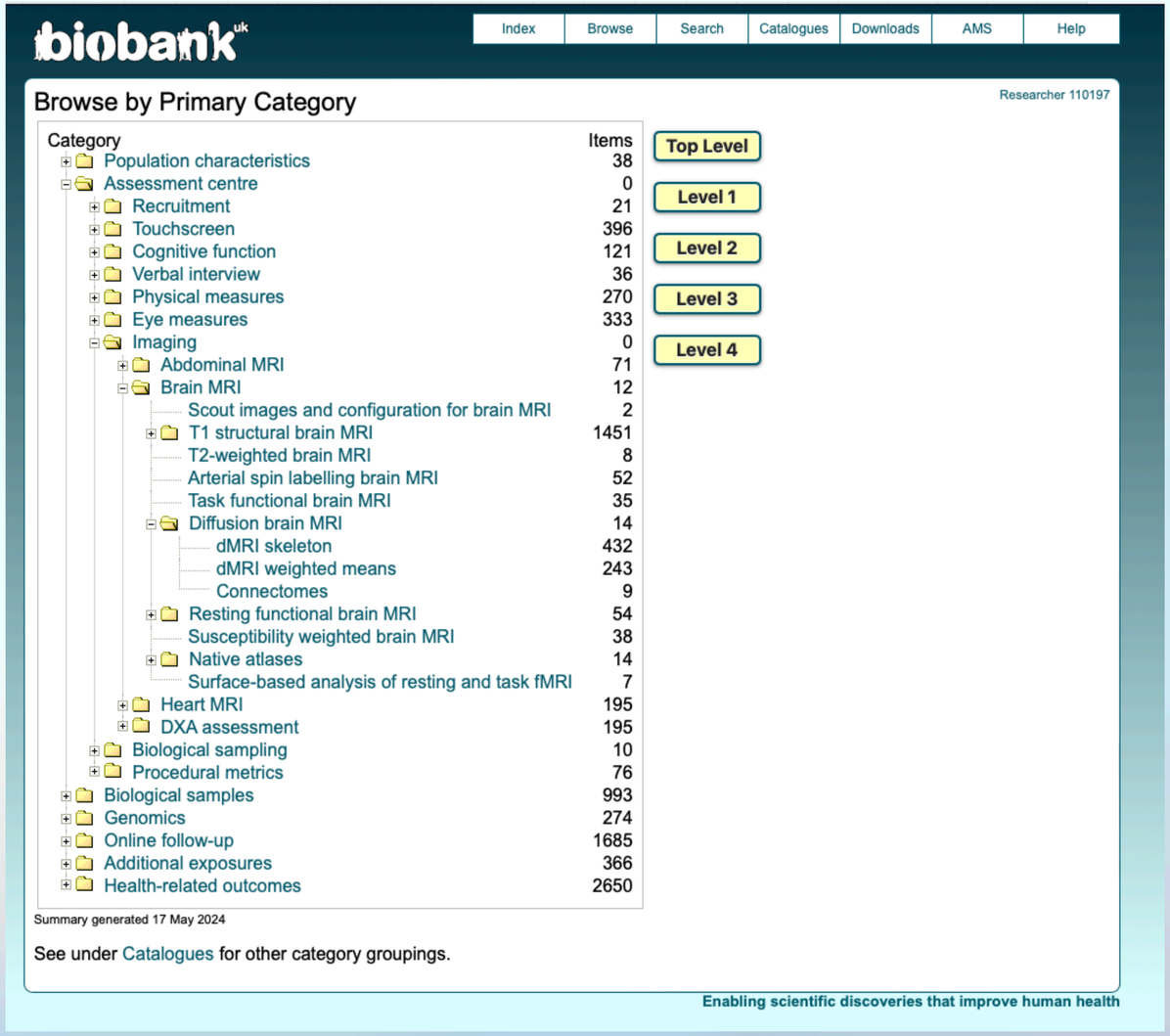




Sources of Health Data

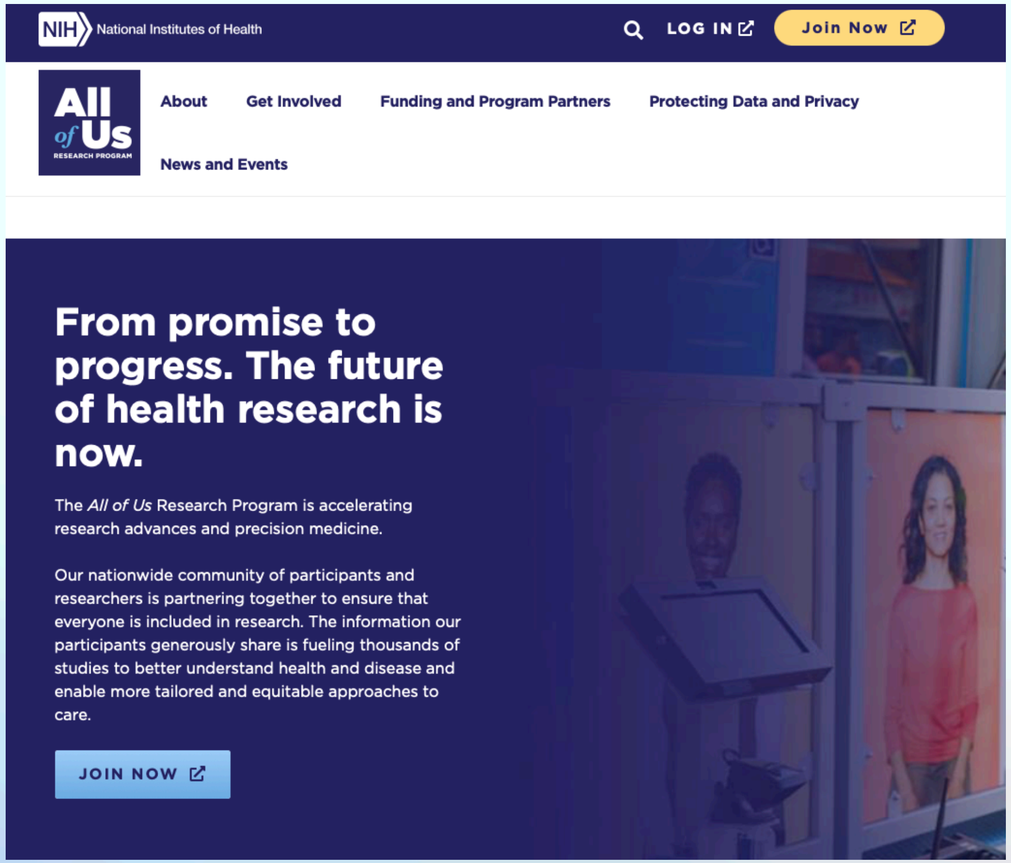
UKBB (UK Biobank) - https://www.ukbiobank.ac.uk/ (500k individuals)

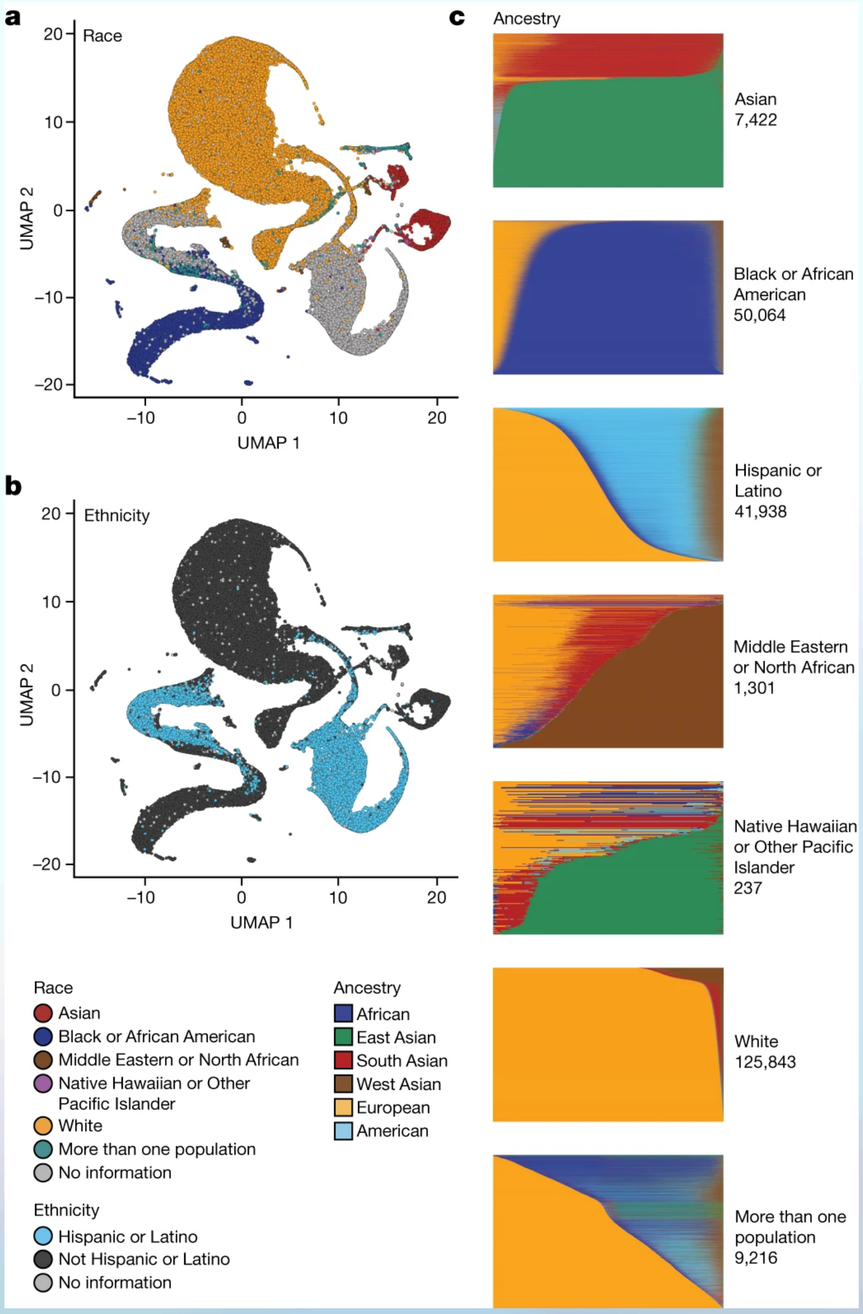




Sources of Health Data

AllOfUs - https://allofus.nih.gov/ (1.3m participants)

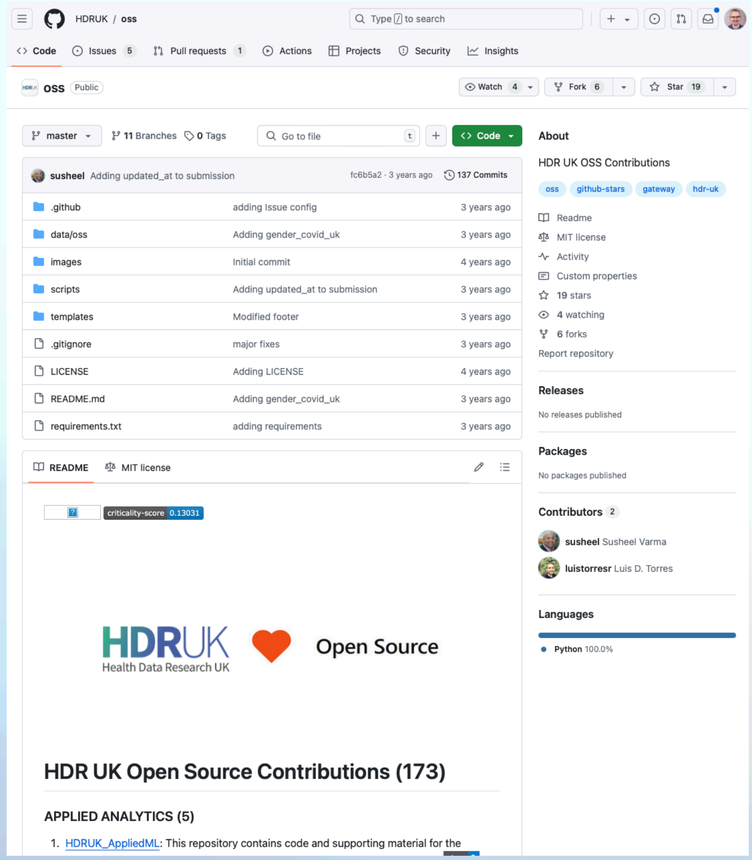


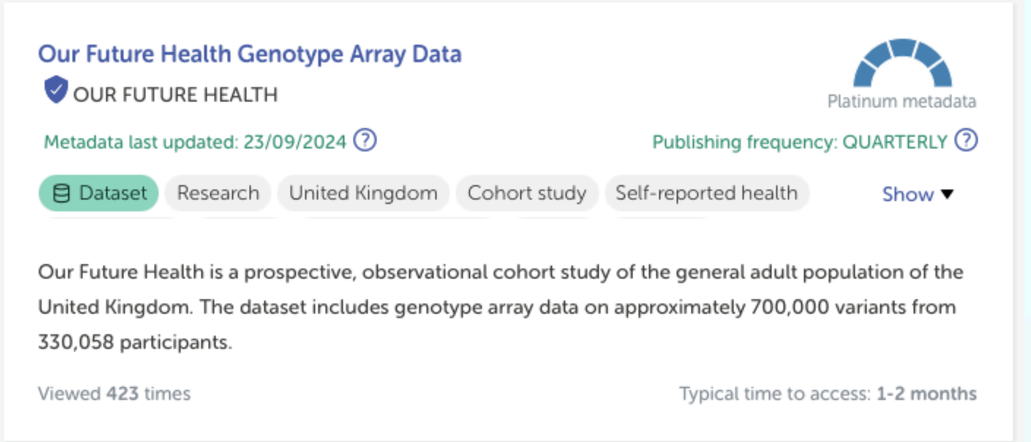


The All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. Nature 627, 340–346 (2024). https://doi.org/10.1038/s41586-023-06957-x

Sources of Health Data

HDR-UK (Health Data Research UK) - https://www.hdruk.ac.uk/





Programmatic Bulk Download

We can use a number of Python approaches to achieve bulk download programmatically rather than having to deal with manual downloading and saving of data:

URL Fetching Libraries

- `requests` - `requests.get(url).content` followed by saving the content to file.
- `urllib` - `urllib.request.urlretrieve(url, filename)` this downloads to file.
- `wget` (wrapper for GNU wget) - `wget.download(url)` this downloads to file.

Threading of multiple downloads simultaneously

- `Scrapy` - large-scale web crawling and bulk data extraction, efficient for downloading data from multiple URLs by defining spiders.
- `pycurl` - curl interface. good for handling multiple simultaneous HTTP/FTP transfers.

Handling Authentication & Session Management

- `requests.Session` to persist certain parameters across requests, can handle login and cookie management for downloading from authenticated sources.
- Selenium for Web Scraping, especially for websites that load with JavaScript, can simulate a browser to fetch and download files and interact with web elements.

Application Programming Interfaces (APIs)

Abstraction - APIs provide a simple interface to more complex underlying systems. They abstract the internal workings exposing only what is necessary for the user.

Standardisation - APIs are designed with standard protocols and rules, ensuring consistency

Security - APIs enforce security, can limit access by requiring authentication, and ensure data integrity and confidentiality through encryption.

Scalability - APIs allow systems to handle increased volumes of data or interactions.

Efficiency - Data exchange through APIs can be optimised to reduce the amount of data that needs to be sent.

Documentation - APIs commonly come with comprehensive documentation that details available endpoints, data formats, and methods.

Versioning - APIs often employ version control, allowing them to introduce new features or deprecate old ones without disrupting existing code developed using them.

Customisation - APIs commonly allow the user to specify parameters associated with their query to tailor data retrieval to they specific needs.

NCBI FTP Bulk Download Example

```python
import urllib.request

def download_file_from_ftp(url, output_filename):
    try:
        # Open the URL
        with urllib.request.urlopen(url) as response:
            # Read the data from the URL
            data = response.read()

            # Writing the data to a file
            with open(output_filename, 'wb') as file:
                file.write(data)

        print("File downloaded successfully!")

    except Exception as e:
        print(f"An error occurred: {e}")

# URL of the file you want to download
url = "https://ftp.ncbi.nlm.nih.gov/gene/DATA/README_ensembl"

# Local file path where you want to save the downloaded file
output_filename = "README_ensembl"

# Calling the download function
download_file_from_ftp(url, output_filename)
```

Python

README_ensembl

```
#tax_id ncbi_release    ncbi_assembly   ensembl_release ensembl_assembly
3486    Humulus lupulus Annotation Release 100  drHumLupu1.1    Rapid 2024-
3517    Alnus glutinosa Annotation Release 100  dhAlnGlut1.1    Rapid 2024-
3691    Populus nigra Annotation Release 100    ddPopNigr1.1    Rapid 2024-
3986    Mercurialis annua Annotation Release 101    ddMerAnnu1.2    Rapid 2
6063    Halichondria panicea Annotation Release 100 odHalPani1.1    Rapid 2
6087    Hydra vulgaris Annotation Release 103   Hydra_105_v3    Rapid 2022-
6105    Actinia tenebrosa Annotation Release 100    ASM960242v1 Rapid 2022-
6148    0       Rapid 2023-10   ASM1229514v1
6454    Haliotis rufescens Annotation Release 101   xgHalRufe1.0.p  Rapid 2
6465    Patella vulgata Annotation Release 100  xgPatVulg1.1    Rapid 2022-
6500    Aplysia californica Annotation Release 102  AplCal3.0   Rapid 2021-
6526    Biomphalaria glabrata Annotation Release 101    xgBioGlab47.1   Rap
6549    Mytilus californianus Annotation Release 100    xbMytCali1.0.p  Rap
6565    Crassostrea virginica Annotation Release 100    C_virginica-3.0 Rap
6573    Mizuhopecten yessoensis Annotation Release 100  ASM211388v2 Rapid 2
6579    Pecten maximus Annotation Release 100   xPecMax1.1  Rapid 2023-08
6596    Mercenaria mercenaria Annotation Release 100    ASM1480567v1.1  Rap
6604    Mya arenaria Annotation Release 100 ASM2691426v1    Rapid 2023-04
6669    Daphnia pulex Annotation Release 100    ASM2113471v1    Rapid 2022-
6687    Penaeus monodon Annotation Release 100  NSTDA_Pmon_1    Rapid NSTDA
6689    Penaeus vannamei Annotation Release 100 ASM378908v1 Rapid 2022-04
```

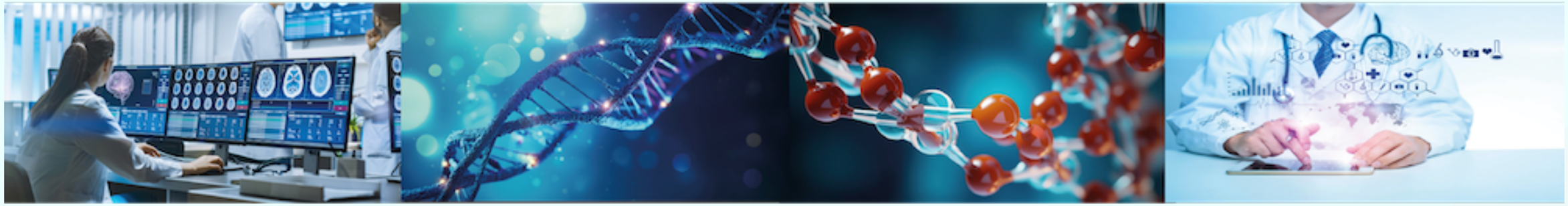
KEGG API Download Example

```python
import requests

def download_kegg_pathway_data(pathway_id):
    # URL for the pathway data
    data_url = f"http://rest.kegg.jp/get/{pathway_id}"

    # Make the HTTP request for the pathway data
    response = requests.get(data_url)

    # Check if the request was successful
    if response.status_code == 200:
        # Write the data content to a file
        with open(f"{pathway_id}.txt", 'w') as file:
            file.write(response.text)
        print(f"Pathway data saved as {pathway_id}.txt")
    else:
        print("Failed to retrieve pathway data. Status code:", response.status_code)

# Example pathway ID for Glycolysis / Gluconeogenesis
pathway_id = "map00010"

# Call the functions to download the pathway image and data
download_kegg_pathway_data(pathway_id)
```

Python

map00010.txt

```
ENTRY       map00010                    Pathway
NAME        Glycolysis / Gluconeogenesis
DESCRIPTION Glycolysis is the process of converting glucose into pyruvate a
CLASS       Metabolism; Carbohydrate metabolism
PATHWAY_MAP map00010  Glycolysis / Gluconeogenesis
MODULE      M00001  Glycolysis (Embden-Meyerhof pathway), glucose => pyruva
            M00002  Glycolysis, core module involving three-carbon compound
            M00003  Gluconeogenesis, oxaloacetate => fructose-6P [PATH:map0
            M00307  Pyruvate oxidation, pyruvate => acetyl-CoA [PATH:map000
DBLINKS     GO: 0006096 0006094
REFERENCE
  AUTHORS   Nishizuka Y (ed).
  TITLE     [Metabolic Maps] (In Japanese)
  JOURNAL   Tokyo Kagaku Dojin (1980)
REFERENCE
  AUTHORS   Nishizuka Y, Seyama Y, Ikai A, Ishimura Y, Kawaguchi A (eds).
  TITLE     [Cellular Functions and Metabolic Maps] (In Japanese)
  JOURNAL   Tokyo Kagaku Dojin (1997)
REFERENCE
  AUTHORS   Michal G.
  TITLE     Biochemical Pathways
  JOURNAL   Wiley (1999)
REL_PATHWAY map00020  Citrate cycle (TCA cycle)
            map00030  Pentose phosphate pathway
            map00500  Starch and sucrose metabolism
            map00620  Pyruvate metabolism
            map00640  Propanoate metabolism
            map00710  Carbon fixation by Calvin cycle
KO_PATHWAY  ko00010
///

```



Programming for Biomedical Informatics

Next Lecture this Thursday - “Finding & Fetching Data”

Please Bring your Laptop!

Ask Questions on the Piazza Discussion Board

https://github.com/biomedical-informatics/pbi

THE UNIVERSITY of EDINBURGH
informatics