

Programming for Biomedical Informatics

Lecture 3 - Introduction to the Biomedical Dataverse

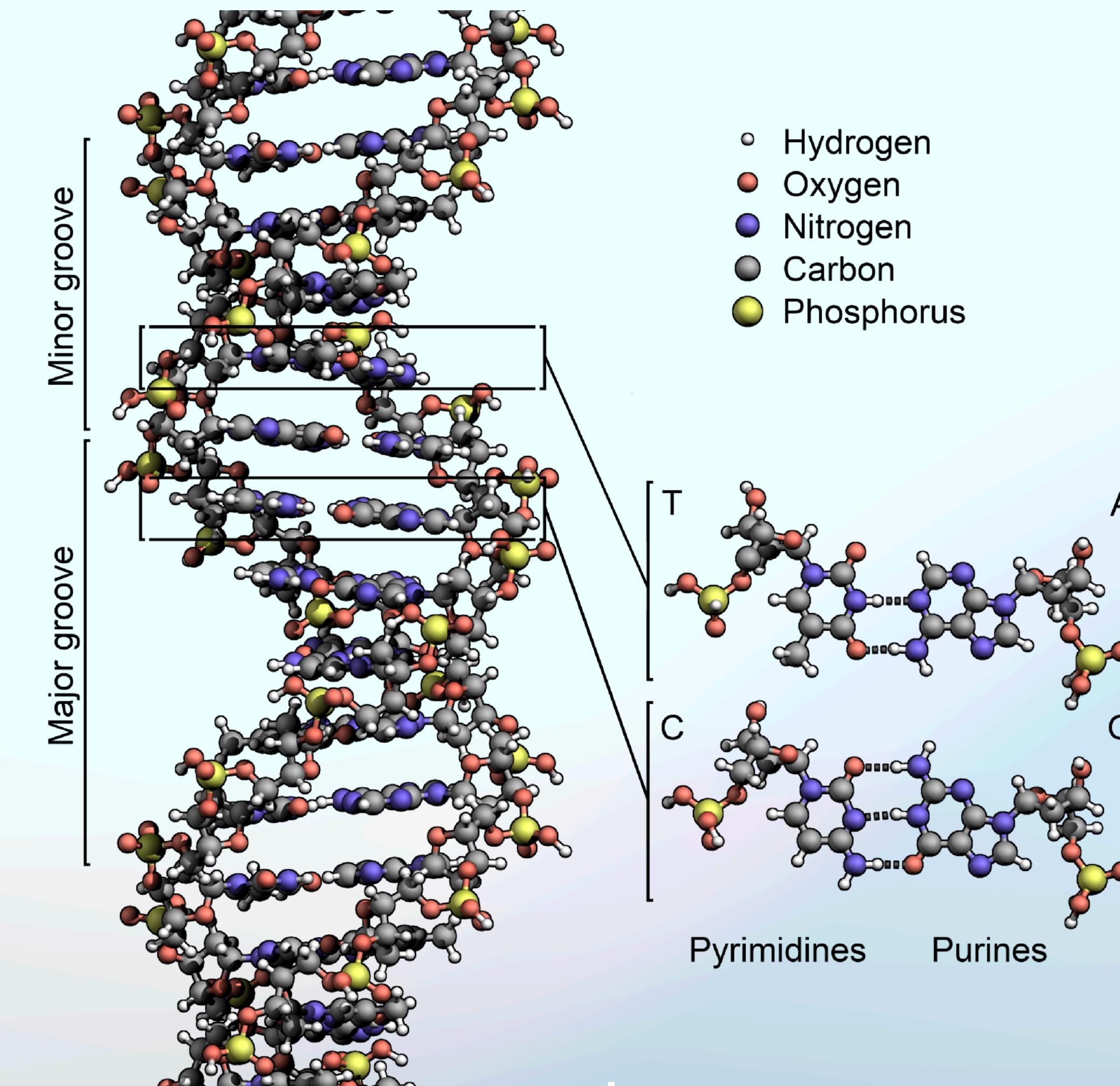
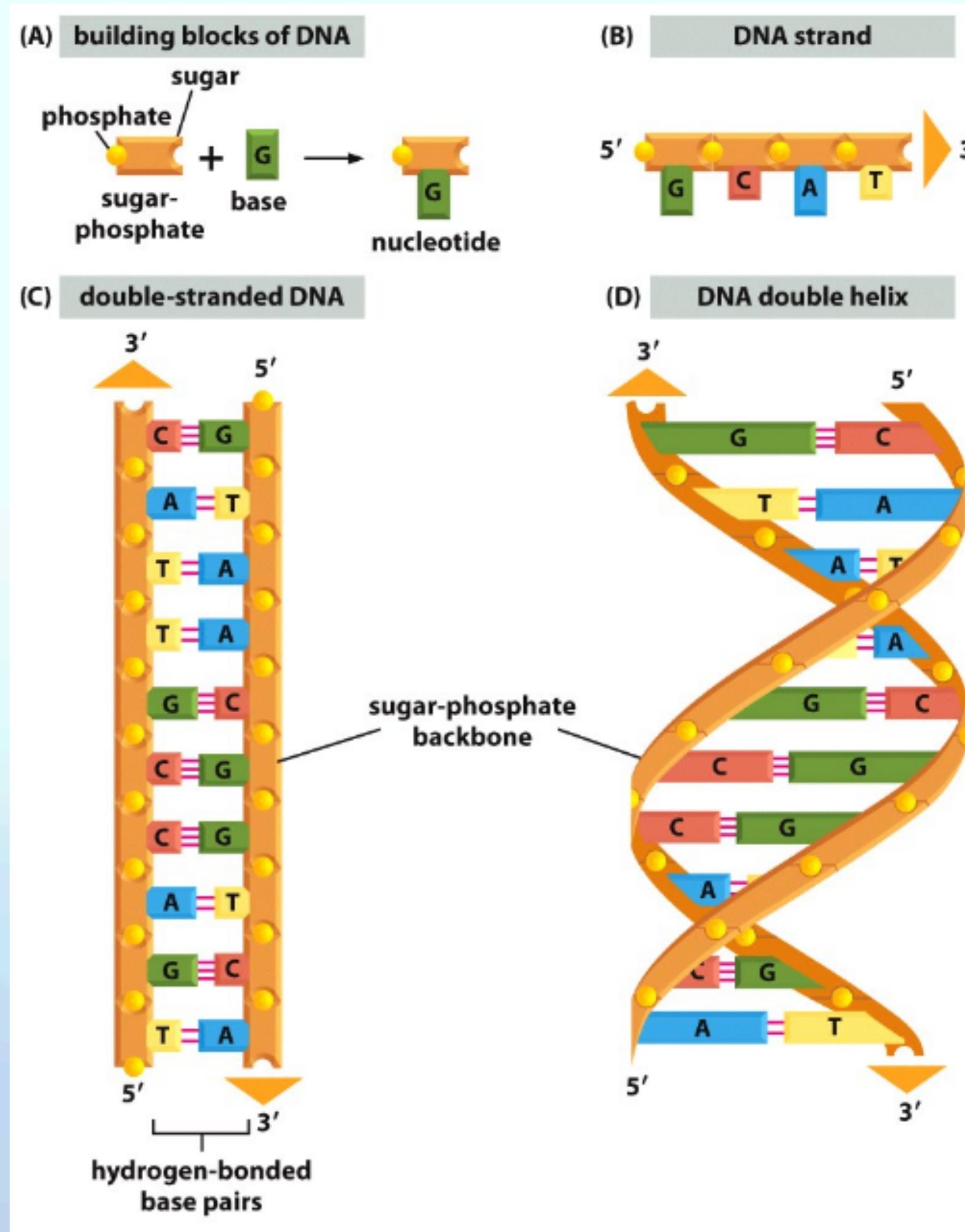
<https://github.com/biomedical-informatics/pbi>

Ian Simpson
ian.simpson@ed.ac.uk

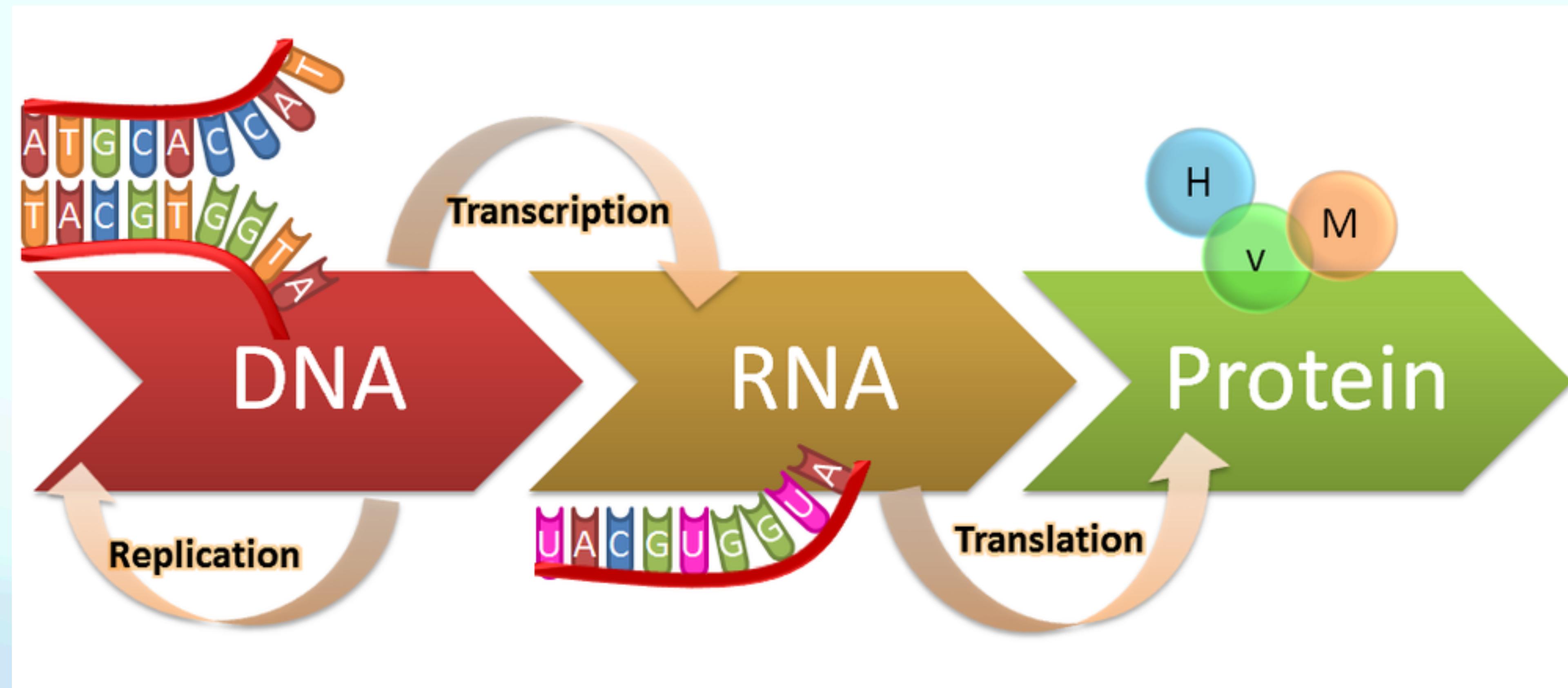
Primer on Central Dogma (optional)



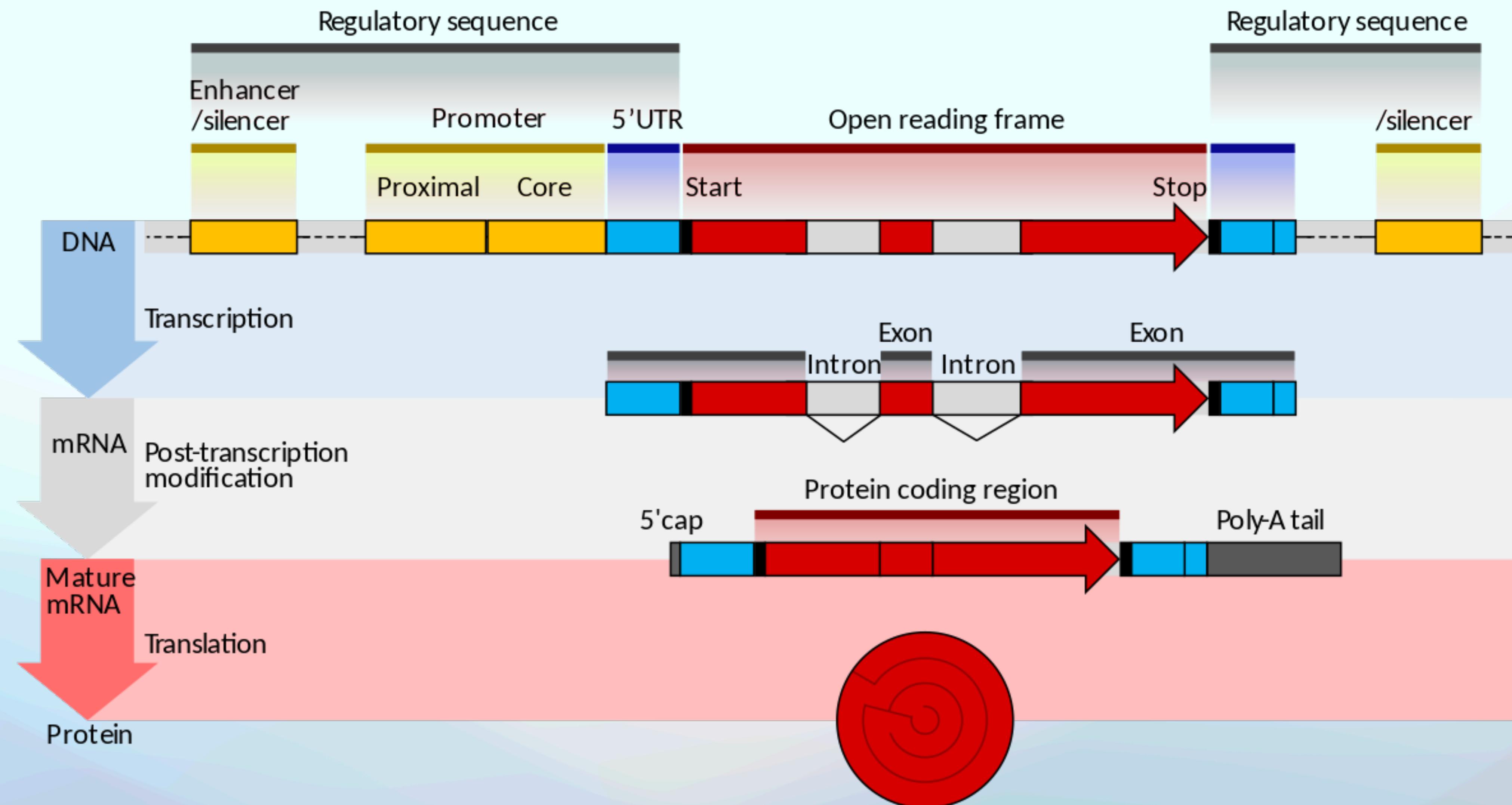
Deoxy-riboNucleic Acid (DNA)



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	TCA		TAA Stop (Ochre)	TGA Stop (Opal)	A							
	TTG	TCG		TAG Stop (Amber)	TGG (Trp/W) Tryptophan	G							
C	CTT (Leu/L) Leucine	CCT	(Pro/P) Proline	CAT (His/H) Histidine	CGT	T							
	CTC	CCC		CAC	CGC	C							
	CTA	CCA		CAA (Gln/Q) Glutamine	CGA	A							
	CTG	CCG		CAG	CGG	G							
A	ATT	ACT	(Thr/T) Threonine	AAT (Asn/N) Asparagine	AGT (Ser/S) Serine	T							
	ATC (Ile/I) Isoleucine	ACC		AAC	AGC	C							
	ATA	ACA		AAA (Lys/K) Lysine	AGA	A							
	ATG ^[A] (Met/M) Methionine	ACG		AAG	AGG (Arg/R) Arginine	G							
G	GTT	GCT	(Ala/A) Alanine	GAT (Asp/D) Aspartic acid	GGT	T							
	GTC	GCC		GAC	GGC	C							
	GTA	GCA		GAA (Glu/E) Glutamic acid	GGA	A							
	GTG	GCG		GAG	GGG	G							

Reading mRNA Sequences

Reading frame #1

5'-AGUCUUACCGCAUUGUGG-3'

Ser--Leu--Thr--Ala--Leu- Trp

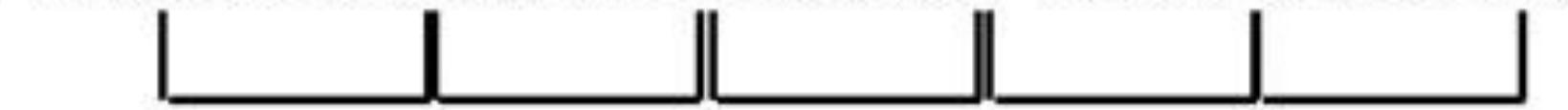
Start: AUG
Stop: UAA, UAG, UGA

Reading frame #2

5'-AGUCUUACCGCAUUGUGG-3'

Val--Leu--Pro--His--Cys

Reading frame #3

5'-AGUCUUACCGCAUUGUGG-3'

Ser--Tyr--Arg--Ile--Val

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 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/>

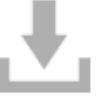
National Library of Medicine
National Center for Biotechnology Information

All Databases

Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit Deposit data or manuscripts into NCBI databases 

Download Transfer NCBI data to your computer 

Learn Find help documents, attend a class or watch a tutorial 

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog
Updated Bacterial and Archaeal Reference Genome Collection now Available! 19 Sep 2024
Download the updated bacterial and archaeal genome collection now!

RefSeq Release 226 is Available! 17 Sep 2024
Check out RefSeq release 226, now available online and from the FTP site. You can access RefSeq data.

NCBI's Read Assembly and Annotation Pipeline Tool (RAPT) to Retire December 2024 16 Sep 2024
As of December 2024, NCBI's rapt tool

[More...](#)

e!Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog Login/Register

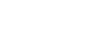
Tools [All tools](#)

BioMart > Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT > Search our genomes for your DNA or protein sequence

Variant Effect Predictor > Analyse your own variants and predict the functional consequences of known and unknown variants

Search All species for
e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

All genomes -- Select a species -- 

Favourite genomes  Human GRCh38.p14
 Pig breeds Pig reference genome and 12 additional breeds
 Mouse GRCm39
 Zebrafish GRCz11

[View full list of all species](#)

NCBI News & Blog
Updated Bacterial and Archaeal Reference Genome Collection now Available! 19 Sep 2024
Download the updated bacterial and archaeal genome collection now!

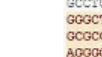
RefSeq Release 226 is Available! 17 Sep 2024
Check out RefSeq release 226, now available online and from the FTP site. You can access RefSeq data.

NCBI's Read Assembly and Annotation Pipeline Tool (RAPT) to Retire December 2024 16 Sep 2024
As of December 2024, NCBI's rapt tool

Compare genes across species 

Find SNPs and other variants for my gene 
GTTATACATT
CCTAAAGCTT
CTTCTAATT
GAACTTCC

Gene expression in different tissues 

Retrieve gene sequence 
GTCTGGCTCGGCTTG
GGCTCTTGGCGGAGAC
GCAGCTCTGTCGCGCT
AAGGGACAGATTGTTG
CACCTCTGACCGGTT
CCCCAGTCCAGCGTGGCG

Find a Data Display 
TABLE
HEATMAP
SEQUENCE
PIE CHART

Use my own data in Ensembl 

Ensembl Rapid Release
New assemblies with gene and protein annotation every two weeks.
Note: species that already exist on this site will continue to be updated with the full range of annotations. Go

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project. Rapid Release news 

Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at [EMBL-EBI](#) and our software and data are freely available. Our [acknowledgements page](#) includes a list of current and previous funding bodies. [How to cite Ensembl](#) in your own publications.

Ensembl release 112 - May 2024 © EMBL-EBI Permanent link - View in archive site

 GLOBAL CORE BIODATA RESOURCE 

Sources of Sequence & Structural Data

UniProt (Unified Proteins) - <https://www.uniprot.org/>

PDB (Protein Data Bank) - <https://www.rcsb.org/>

UniProt BLAST Align Peptide search ID mapping SPARQL

Release 2024_04 | Statistics 📈 🏷️ 📩 Help

UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. [Cite UniProt™](#)

Proteins UniProt Knowledgebase
Reviewed (Swiss-Prot) 571,864
Unreviewed (TrEMBL) 245,324,902

Species Proteomes

Protein Clusters UniRef
Clusters of protein sequences at 100%, 90% & 50% identity

Sequence archive UniParc
Non-redundant archive of publicly available protein sequences seen across different databases

Supporting Data

- Taxonomy
- Keywords
- Literature Citations
- Human diseases
- Cross-referenced databases
- Subcellular locations
- Automatic annotations: UniRule & ARBA

Liquid yellow protein spotlight
When the opportunity to write a piece on urine arose, I thought "wonderful, here's something we can all relate to". I had no idea, however, where it was going to lead me: ...

#UsingUniProt - DisCa... In recent years a wealth of information has become available about genetic variations that...

How artificial intellige... A conversation with machine learning engineer Andreea Gane. At UniProt we are very...

Latest News

View release archive

Forthcoming changes
Planned changes for UniProt
[UniProt release 2024_04](#)
Oocyte waste disposal strategy: 'store to degrade later' | Removal of the cross-references to CLAE...
[UniProt release 2024_03](#)
The culprit for extreme morning sickness identified | Removal of the cross-references to...

UniProt release 2024_02

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB 225,158 Structures from the PDB 1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entry ID(s), or sequence Include CSM Advanced Search | Browse Annotations Help

PDB-101 OPDB EMDDataResource NAKB wwwPDB Foundation PDB-Dev

Access Computed Structure Models (CSMs) of available model organisms Learn more

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

Experimentally-determined 3D structures from the Protein Data Bank (PDB) archive

Computed Structure Models (CSM) from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

Explore NEW Features PDB-101 Training Resources

September Molecule of the Month

Carbon Capture Mechanisms

Sources of Imaging Data

Allen Brain Atlas - <https://www.brain-map.org/>

Human Protein Atlas - <https://www.proteinatlas.org/>

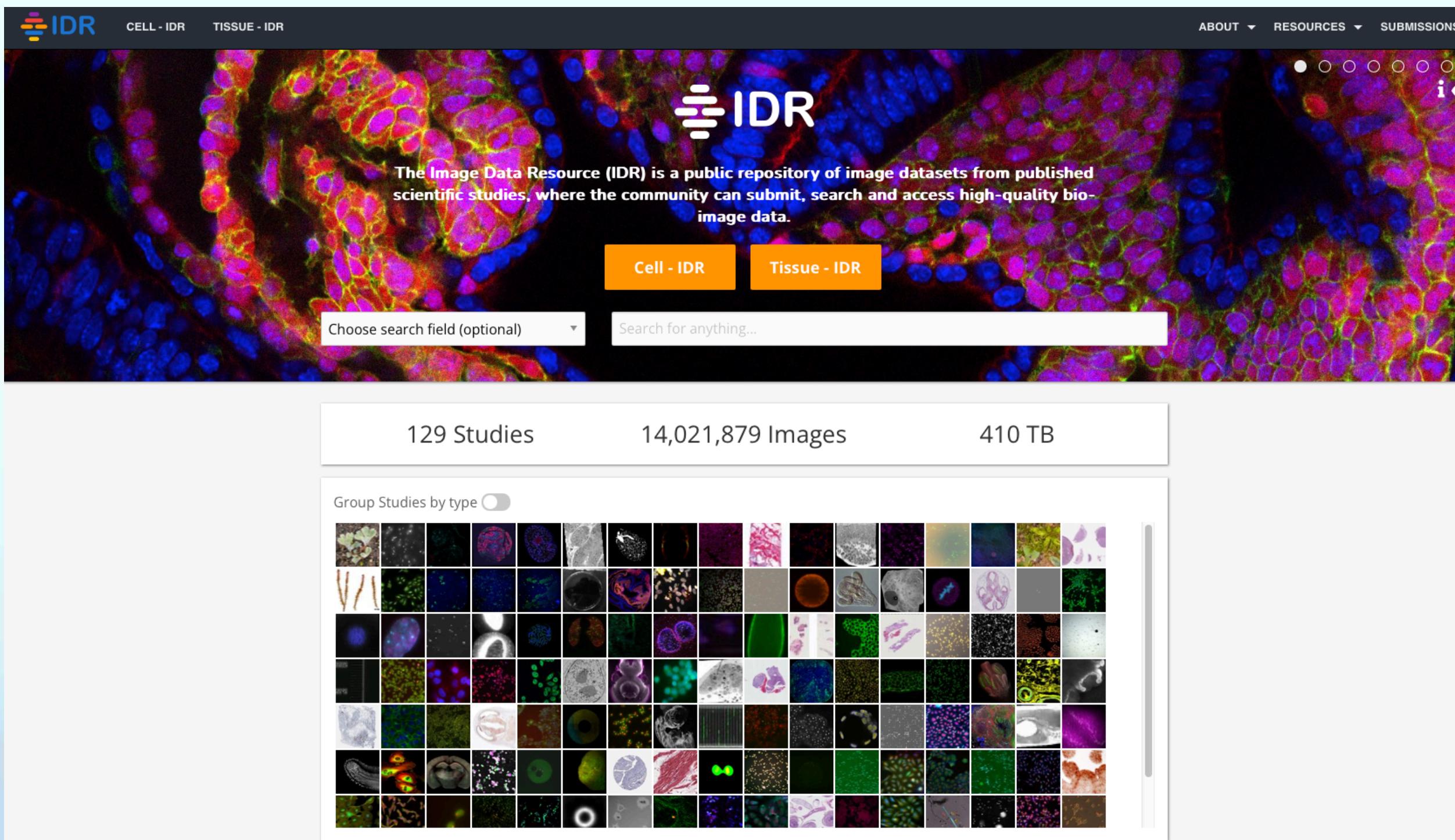
The Allen Brain Map website features a dark blue header with navigation links for "ALLEN BRAIN MAP", "Our Research", "Atlases and Data", and "Technical Resources". Below the header, a large banner displays several circular brain imaging slices in various colors (blue, green, red). A central text box reads: "Accelerating progress toward understanding the brain". Below the banner, a section discusses open science practices, mentioning the release of datasets and software. It includes a quote from Emre Mazy: "Almost all the research I've done for the last seven years would have been impossible without these data." There are sections for "Anatomy", "Cell Types", "Circuits & Behavior", and "Connectivity", each with a circular icon and a brief description. At the bottom, there are "Explore" buttons for each category.

The Human Protein Atlas website has a light blue header with the title "THE HUMAN PROTEIN ATLAS" and a logo consisting of three colored hexagons (purple, orange, yellow). Below the header, a search bar is present with placeholder text "e.g. ACE2, GFAP, EGFR" and buttons for "Search" and "Fields". The main content area features a "Research Articles" section with a "Read our key publications" button and a grid of journal covers from Science magazine. To the right, there are sections for "Fernström award to HPA affiliated researcher" (mentioning Dr. Cecilia Lindskog) and "The Human Protein Atlas is the GCBR of the Week" (mentioning the Global Core Biodata Resource). The bottom half of the page is divided into a grid of 14 colored boxes, each representing a different aspect of protein research: TISSUE, BRAIN, SINGLE CELL TYPE, TISSUE CELL TYPE, PATHOLOGY, DISEASE, IMMUNE CELL, BLOOD PROTEIN, SUBCELLULAR, CELL LINE, STRUCTURE, and INTERACTION. Each box contains a small image and a brief description.

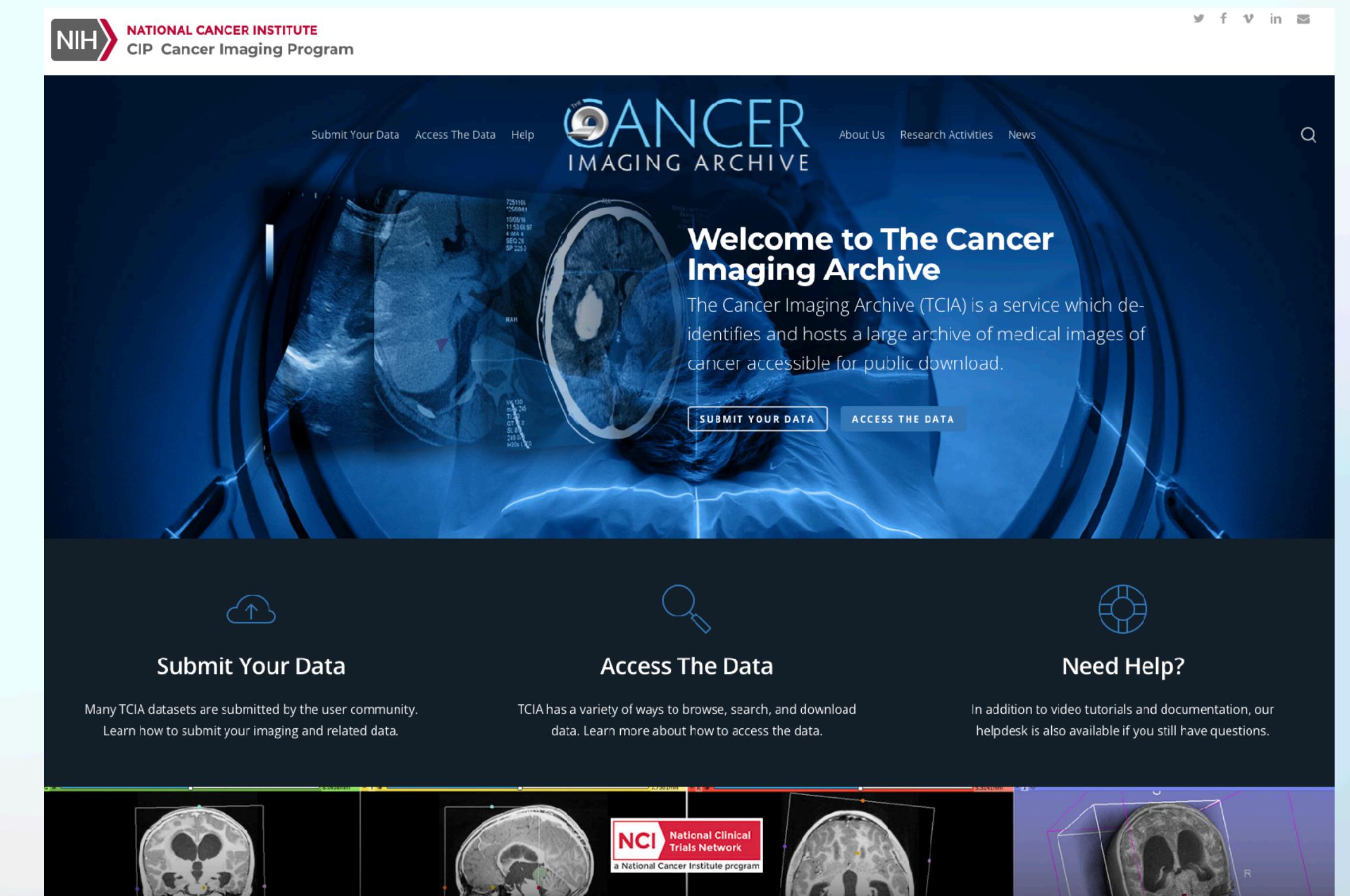
Sources of Imaging Data

Image Data Resource - <https://idr.openmicroscopy.org/>

The Cancer Imaging Archive - <https://www.cancerimagingarchive.net/>



The screenshot shows the IDR homepage. At the top, there's a large image of a tissue sample stained with various colors (red, green, blue). The IDR logo is in the center. Below the image, a text box reads: "The Image Data Resource (IDR) is a public repository of image datasets from published scientific studies, where the community can submit, search and access high-quality bio-image data." There are two orange buttons: "Cell - IDR" and "Tissue - IDR". Below these buttons is a search bar with the placeholder "Search for anything...". Underneath the search bar, it says "Choose search field (optional)" with a dropdown menu. At the bottom of the main section, there are three statistics: "129 Studies", "14,021,879 Images", and "410 TB". Below this, there's a grid of small thumbnail images representing different datasets.



The screenshot shows the TCIA homepage. At the top, there's a banner with the NIH logo and "NATIONAL CANCER INSTITUTE CIP Cancer Imaging Program". Below the banner, the TCIA logo is prominently displayed. The main heading is "Welcome to The Cancer Imaging Archive". A subtext explains: "The Cancer Imaging Archive (TCIA) is a service which identifies and hosts a large archive of medical images of cancer accessible for public download." There are two buttons: "SUBMIT YOUR DATA" and "ACCESS THE DATA". Below the main heading, there are three sections: "Submit Your Data" (with an upload icon), "Access The Data" (with a magnifying glass icon), and "Need Help?" (with a circular icon). At the bottom, there's a footer with the NCI logo and "National Clinical Trials Network a National Cancer Institute program".

Sources of Biomedical Reference Data

PubMed - <https://pubmed.ncbi.nlm.nih.gov/>

OMIM (Online Mendelian Inheritance in Man) - <https://www.omim.org/>

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information

PubMed®

Search

Advanced

PubMed® comprises more than 37 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

Learn

About PubMed
FAQs & User Guide
Finding Full Text

Find

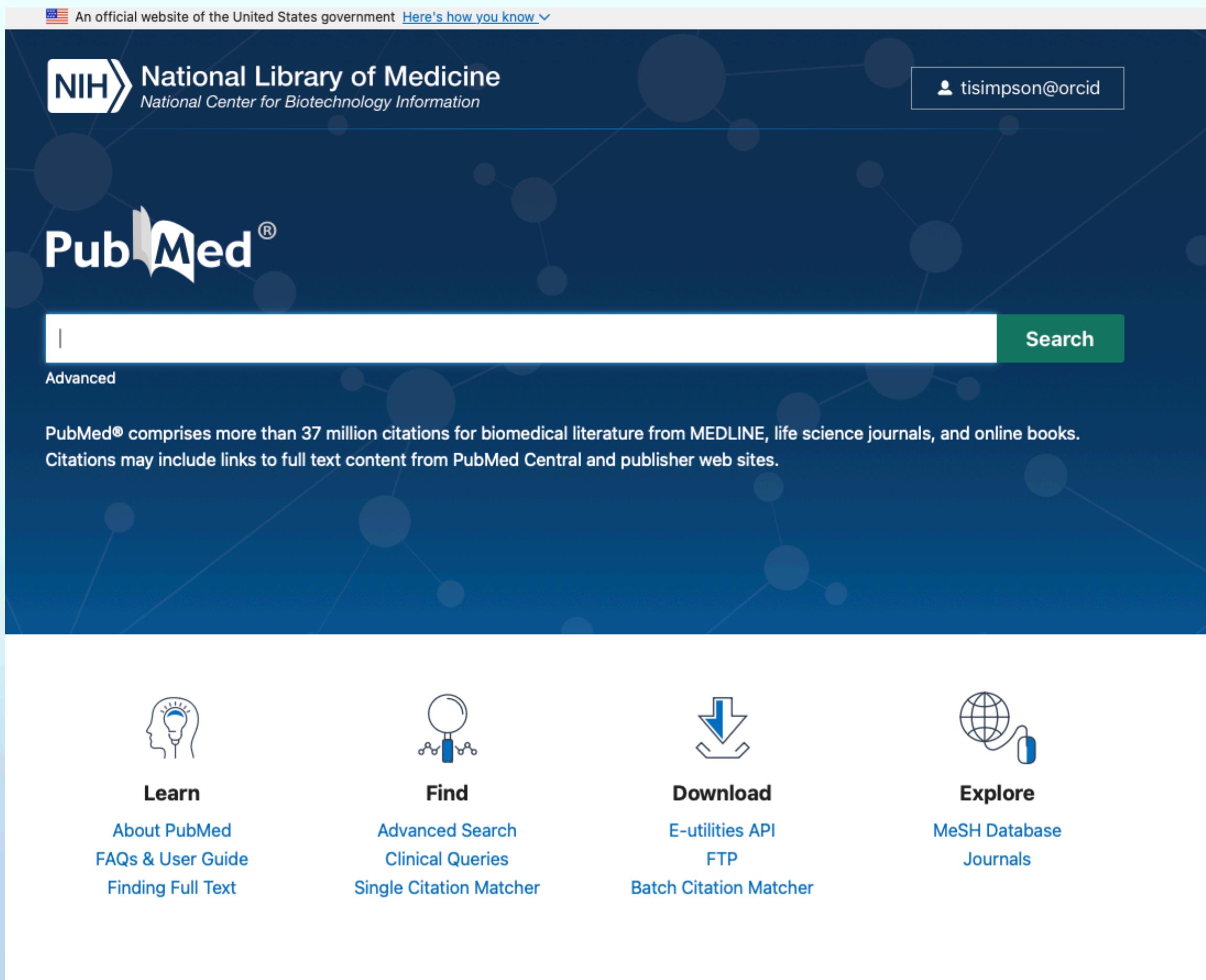
Advanced Search
Clinical Queries
Single Citation Matcher

Download

E-utilities API
FTP
Batch Citation Matcher

Explore

MeSH Database
Journals



OMIM 50 YEARS Human Genetics Knowledge for the World

OMIM®

An Online Catalog of Human Genes and Genetic Disorders

Updated September 20th, 2024

Search OMIM for clinical features, phenotypes, genes, and more... 

Advanced Search : OMIM, Clinical Synopses, Gene Map
Need help? : Example Searches, OMIM Search Help,  OMIM Video Tutorials
Mirror site : <https://mirror.omim.org>

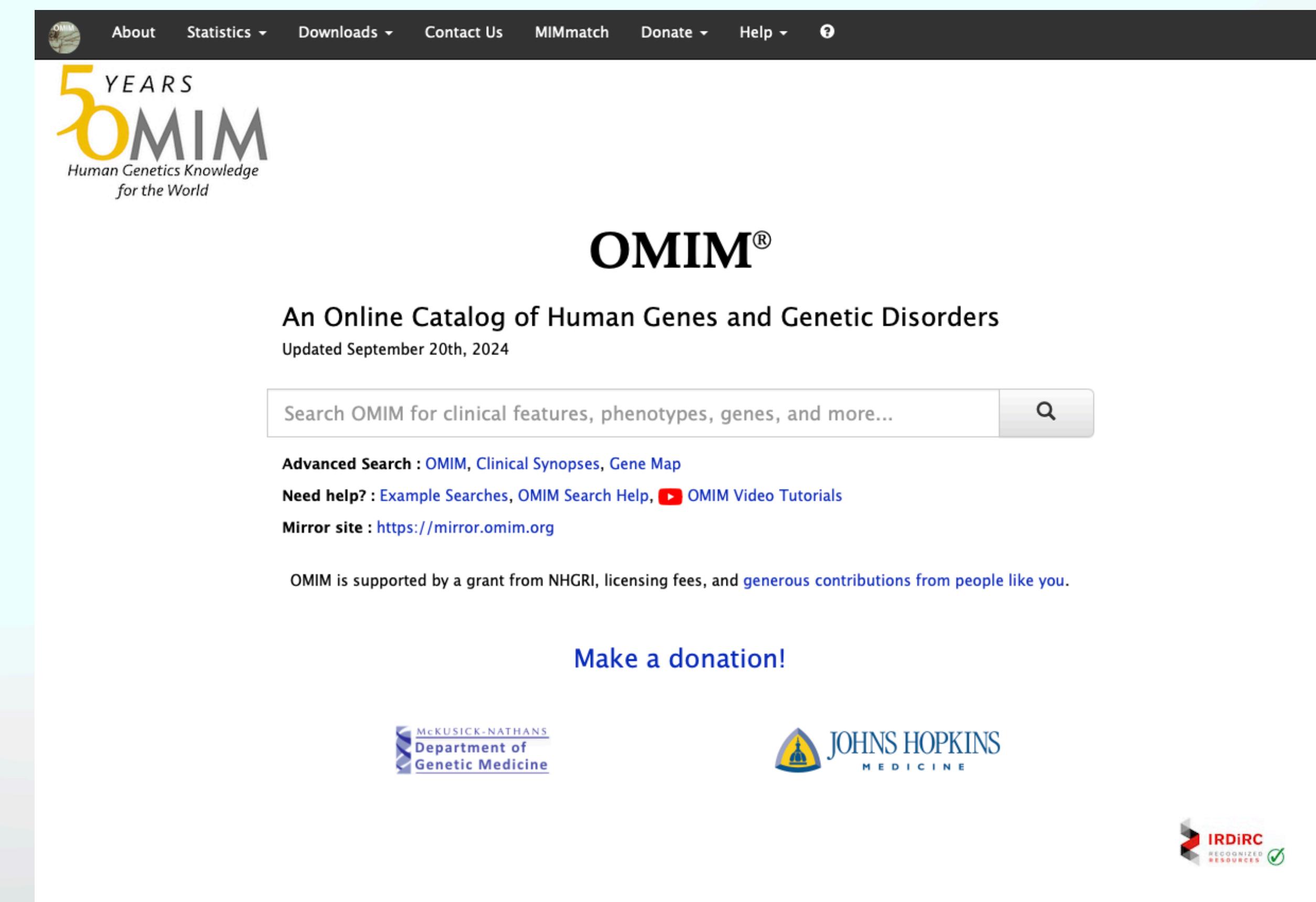
OMIM is supported by a grant from NHGRI, licensing fees, and [generous contributions from people like you](#).

Make a donation!

McKUSICK-NATHANS Department of Genetic Medicine

JOHNS HOPKINS MEDICINE

IRDIRC RECOGNIZED RESOURCES



Sources of Biomedical Reference Data

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

BioPortal Ontologies Search Annotator Recommender Mappings Login Support ▾

Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies

Search for a class
Enter a class, e.g. Melanoma Advanced search

Find an ontology
Start typing ontology name, then choose from list Browse ontologies ▾

Ontology visits (August 2024)

Ontology	Visits (August 2024)
MEDDRA	~32,000
RXNORM	~15,000
SNOMEDCT	~15,000
NDDF	~4,000
RCD	~1,000

Statistics

Category	Value
Ontologies	1,147
Classes	15,663,490
Properties	36,286
Mappings	99,548,310

[More](#)

reactome About ▾ Content ▾ Docs ▾ Tools ▾ Community ▾ Download

e.g. O95631, NTN1, signaling by EGFR, glucose Go!

Home > Docs > Userguide

Userguide

- Developer's Zone
- Icon Info
- Data Model
- Computationally inferred events
- FAQ
- Linking to Us
- Citing us

What is Reactome?

Reactome is a curated database of pathways and reactions in human biology. Reactions can be considered as pathway 'steps'. Reactome defines a 'reaction' as any event in biology that changes the state of a biological molecule. Binding, activation, translocation, degradation and classical biochemical events involving a catalyst are all reactions. Information in the database is authored by expert biologists, entered and maintained by Reactome's team of Curators and Editorial staff. Reactome content frequently cross-references other resources e.g. NCBI, Ensembl, UniProt, KEGG (Gene and Compound), ChEBI, PubMed and GO. Inferred orthologous reactions Inferred orthologous reactions are available for 15 non-human species including mouse, rat, chicken, puffer fish, worm, fly, yeast, rice, and Arabidopsis

What is this guide For?

This guide introduces features of the Reactome website using a combination of short explanations and exercises. You will learn how to search, interpret the views, use the tools and if necessary find documentation or contact us for help.

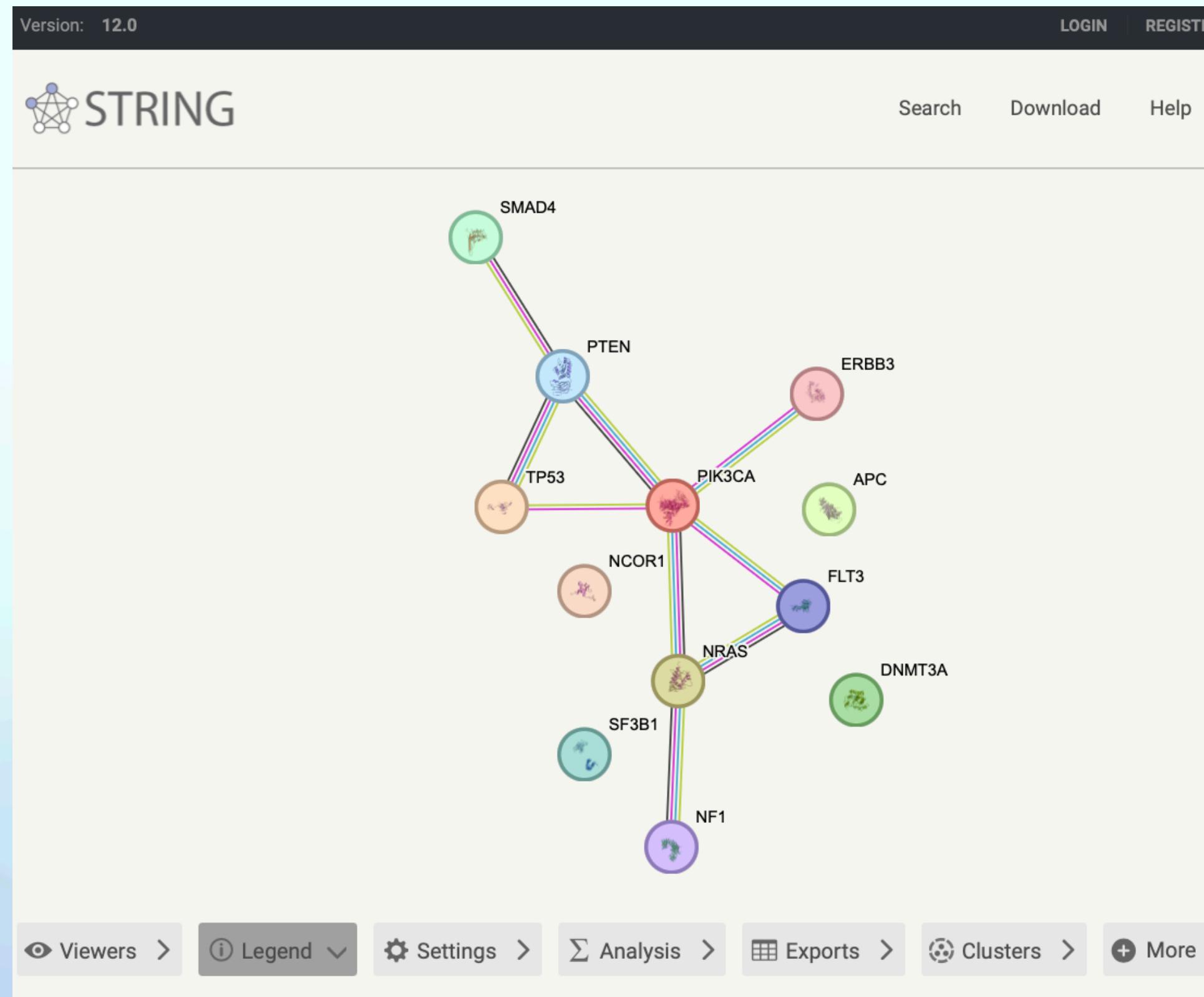
Pathway Browser **Details Panel** **How do I search?** **Analysis Tools**

Cytomics **Diseases** **ReactomeFlViz**

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)



SNOMED CT Browser Release: International Edition Version: 2024-09-01 Perspective: Full Feedback About

Taxonomy Search Favorites Refset

Search Type at least 3 characters ✓ Example: shou fra

Options Search: Prefix any order ▾ Status: Active concepts only ▾ Description type: All ▾ Language Refsets ▾ Group by concept Filter results by Language english 101 Filter results by Semantic Tag assessment scale 3 disorder 88 environment 1 observable entity 2 procedure 4 situation 4

Concept Details Expression Constraint Queries

Concept Details Summary Details Diagram Expression Refsets Members History References Stated Inferred

Parents > Disorder of psychological development (disorder)

Pervasive developmental disorder (disorder) SCTID: 35919005 Pathological process → Pathological developmental process

35919005 | Pervasive developmental disorder (disorder) | en Pervasive developmental disorder (disorder) en Autism spectrum disorder en Autism en Pervasive developmental disorder

Children (39)

- 1p21.3 microdeletion syndrome (disorder)
- Active but odd autism (disorder)
- Activity dependent neuroprotector homeobox related multiple congenital anomalies, intellectual disability, autism spectrum disorder (disorder)
 - > Asperger's disorder (disorder)
 - Atypical autism (disorder)
 - Atypical Rett syndrome (disorder)
 - Autism spectrum disorder due to AUTS2 activator of transcription and developmental regulator deficiency (disorder)
 - Autism spectrum disorder, epilepsy, arthrogryposis syndrome (disorder)

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/>

The screenshot shows the GEO homepage with a blue header bar containing the NCBI logo, Resources, How To, Sign in to NCBI, and a search bar. Below the header, there's a main title "Gene Expression Omnibus" and a brief description of what it is. On the left, there's a sidebar titled "Getting Started" with links to Overview, FAQ, About GEO DataSets, About GEO Profiles, About GEO2R Analysis, How to Construct a Query, and How to Download Data. In the center, there are three main sections: "Tools" (with links to Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, Search GEO Documentation, Analyze a Study with GEO2R, Studies with Genome Data Viewer Tracks, Programmatic Access, FTP Site, and ENCODE Data Listings and Tracks), "Browse Content" (with a Repository Browser showing DataSets: 4348, Series: 236587, Platforms: 26550, and Samples: 7409881), and a "GEO" logo with the text "Gene Expression Omnibus".

The screenshot shows the ENCODE Portal homepage with a dark header bar containing ENCODE, Data, Encyclopedia, Materials & Methods, Help, and a shopping cart icon. Below the header is a search bar with options for "ENCODE" and "SCREEN". The main content area features a large image of a DNA double helix and a grid of 12 cards representing different ENCODE projects: Functional genomics (with a DNA helix icon), Functional characterization (with a pair of scissors icon), Encyclopedia of elements (with a book icon), Rush Alzheimer's (with a brain icon), EN-TEx (with an ENCODE GTEX logo), Deeply profiled cell lines (with a petri dish icon), Protein knockdown (Degron) (with a cartoon protein icon), Computational and integrative products (with a brain and monitor icon), Human donors (with a group of people icon), ENCORE (with an RNA elements logo), Stem cell differentiation (with a stem cell cluster icon), and Imputed experiments (with a lab setup icon).

Sources of Biomedical Experimental Data

HCA (Human Cell Atlas) - <https://www.humancellatlas.org>

GTEx Portal (Adult Genotype Expression Project) - <https://gtexportal.org/>

HUMAN CELL ATLAS DATA PORTAL

Datasets HCA BioNetworks Guides Metadata APIs Updates Follow HCA Search

Explore the datasets of the Human Cell Atlas

Community generated, multi-omic, open data

60.4M Cells 9.0k Donors 462 Projects 774 Labs

Explore Data

HCA Biological Network Atlases

Adipose	Gut	Lung	Pancreas
Breast	Heart	Musculoskeletal	Reproduction
Development	Immune	Nervous System	Skin
Eye	Kidney	Oral and Craniofacial	
Genetic Diversity	Liver	Organoid	

GTEx Portal

About Adult GTEx Publications Access Biospecimens FAQs Contact

Home Downloads Expression Single Cell QTL IGV Browser Tissues & Histology Documentation About

Search Gene or SNP ID

The Genotype-Tissue Expression (GTEx) Portal is a comprehensive public resource for researchers studying tissue and cell-specific gene expression and regulation across individuals, development, and species, with data from 3 NIH projects.

GTEx

The Adult GTEx project is a comprehensive resource of WGS, RNA-Seq, and QTL data from samples collected from 54 non-diseased tissue sites across ~1000 adult individuals.

Explore »

dGTEx

The Developmental GTEx (dGTEx) project is a new effort to study development-specific genetic effects on gene expression and to establish a new data analysis and tissue biobank resource.

*Data Not Yet Available Explore »

NHP-dGTEx

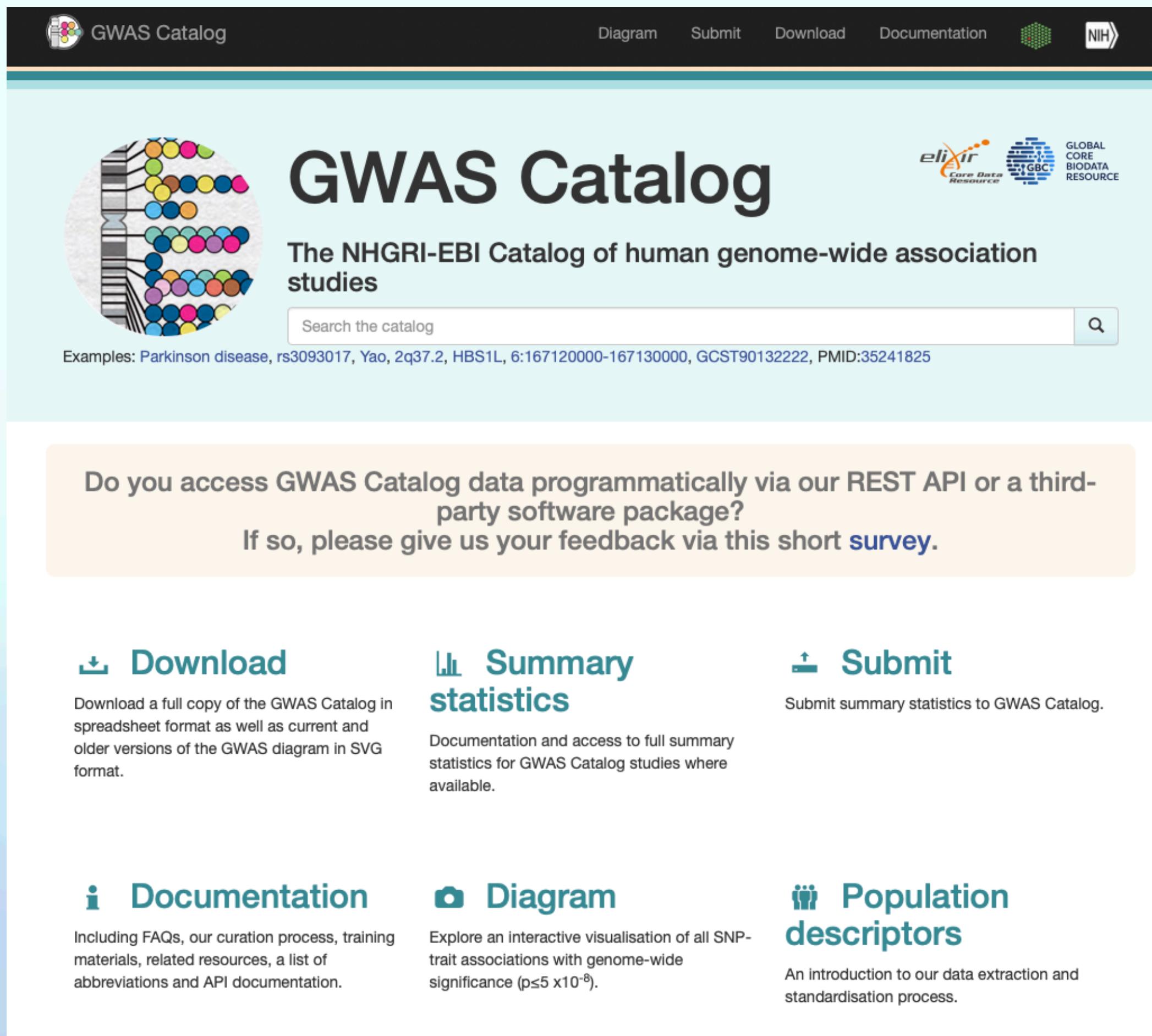
The Non-Human Primate Developmental GTEx (NHP-dGTEx) project is a complement to dGTEx in 2 translational non-human primate model species: the rhesus macaque and common marmoset.

*Data Not Yet Available Explore »

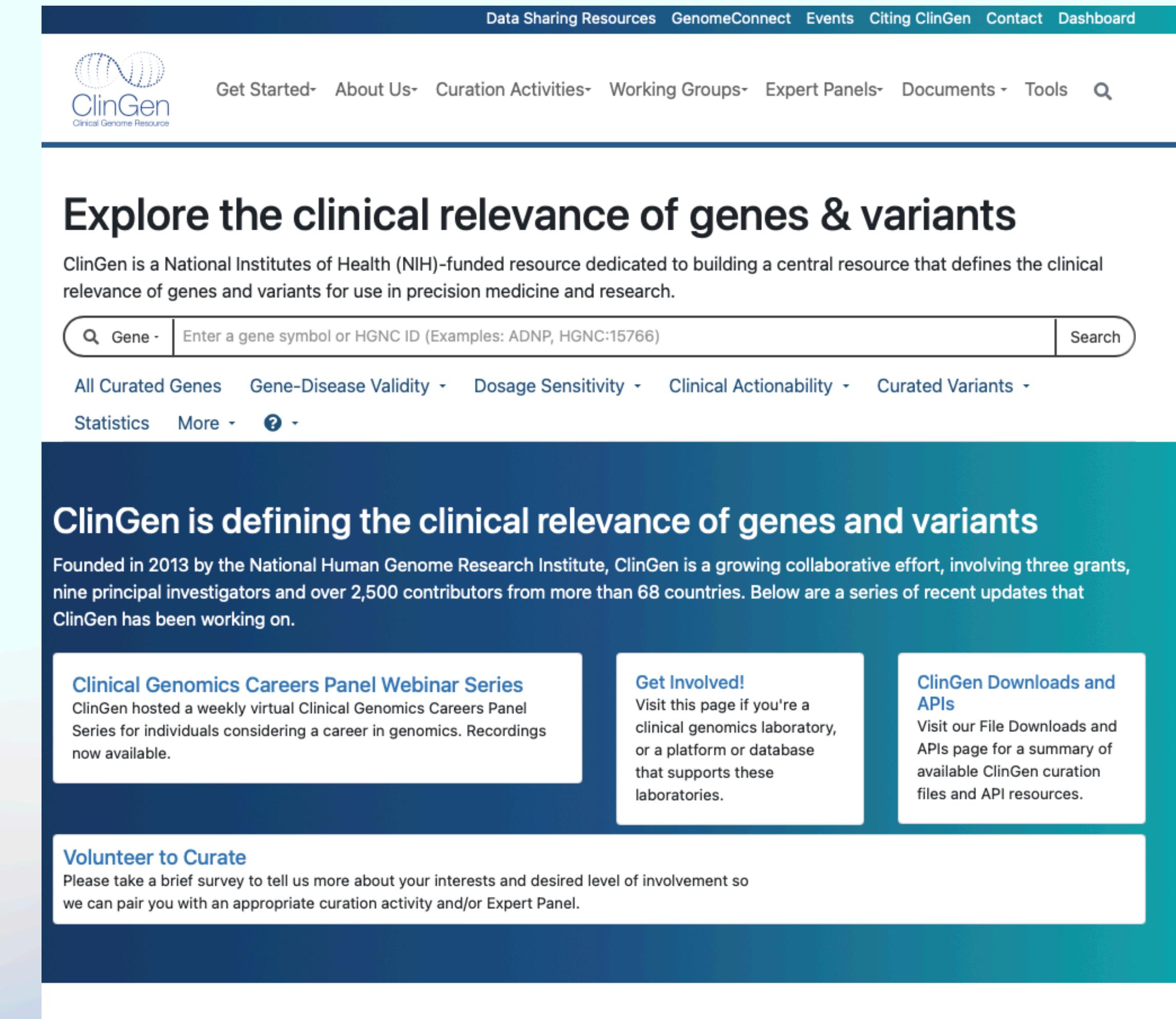
Sources of Biomedical Experimental Data

GWAS Catalog (Genome Wide Association Studies) - <https://www.ebi.ac.uk/gwas/>

ClinGen (The Clinical Genome - variation/disease) - <https://www.clinicalgenome.org/>



The screenshot shows the homepage of the GWAS Catalog. At the top, there is a navigation bar with links for "Diagram", "Submit", "Download", "Documentation", and the NIH logo. Below the navigation bar, there is a search bar with the placeholder "Search the catalog". To the left of the search bar is a circular graphic representing a genome. To the right of the search bar are logos for ELIXIR and the Global Core Biodata Resource. The main title "GWAS Catalog" is prominently displayed in large, bold letters. Below the title, the subtitle "The NHGRI-EBI Catalog of human genome-wide association studies" is shown. A callout box in the center of the page asks if users access the catalog programmatically via a REST API or a third-party software package, with a link to a survey. At the bottom of the page, there are six sections: "Download", "Summary statistics", "Submit", "Documentation", "Diagram", and "Population descriptors". Each section has a brief description and a small icon.



The screenshot shows the homepage of the ClinGen website. At the top, there is a navigation bar with links for "Data Sharing Resources", "GenomeConnect", "Events", "Citing ClinGen", "Contact", and "Dashboard". Below the navigation bar, there is a search bar with the placeholder "Enter a gene symbol or HGNC ID (Examples: ADNP, HGNC:15766)". To the left of the search bar is the ClinGen logo. The main title "Explore the clinical relevance of genes & variants" is displayed in large, bold letters. Below the title, a brief description states that ClinGen is a National Institutes of Health (NIH)-funded resource dedicated to building a central resource that defines the clinical relevance of genes and variants for use in precision medicine and research. A callout box in the center of the page highlights recent updates from the Clinical Genomics Careers Panel Webinar Series. At the bottom of the page, there are three boxes: "Get Involved!", "ClinGen Downloads and APIs", and "Volunteer to Curate".

Sources of Health Data

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

The world's most important health research database

Data drives discovery. We have curated a uniquely powerful biomedical database that can be accessed globally by approved researchers. Use our secure cloud-based platform to explore de-identified data from half a million UK Biobank participants and enable new discoveries to improve public health.

About our data About us

Enable your research Explore your participation Learn more about UK Biobank

Change in how to access data UK Biobank Conference: register today Read our participant newsletter

biobank^{uk}

Index Browse Search Catalogues Downloads AMS Help

Researcher 110197

Browse by Primary Category

Category	Items	
+ Population characteristics	38	Top Level
- Assessment centre	0	
+ Recruitment	21	Level 1
+ Touchscreen	396	
+ Cognitive function	121	
+ Verbal interview	36	
+ Physical measures	270	
+ Eye measures	333	
- Imaging	0	
+ Abdominal MRI	71	Level 2
+ Brain MRI	12	
Scout images and configuration for brain MRI	2	
+ T1 structural brain MRI	1451	Level 3
+ T2-weighted brain MRI	8	
+ Arterial spin labelling brain MRI	52	
+ Task functional brain MRI	35	
+ Diffusion brain MRI	14	
dMRI skeleton	432	Level 4
dMRI weighted means	243	
Connectomes	9	
+ Resting functional brain MRI	54	
+ Susceptibility weighted brain MRI	38	
+ Native atlases	14	
Surface-based analysis of resting and task fMRI	7	
+ Heart MRI	195	
+ DXA assessment	195	
+ Biological sampling	10	
+ Procedural metrics	76	
+ Genomics	993	
+ Online follow-up	274	
+ Additional exposures	1685	
+ Health-related outcomes	366	
	2650	

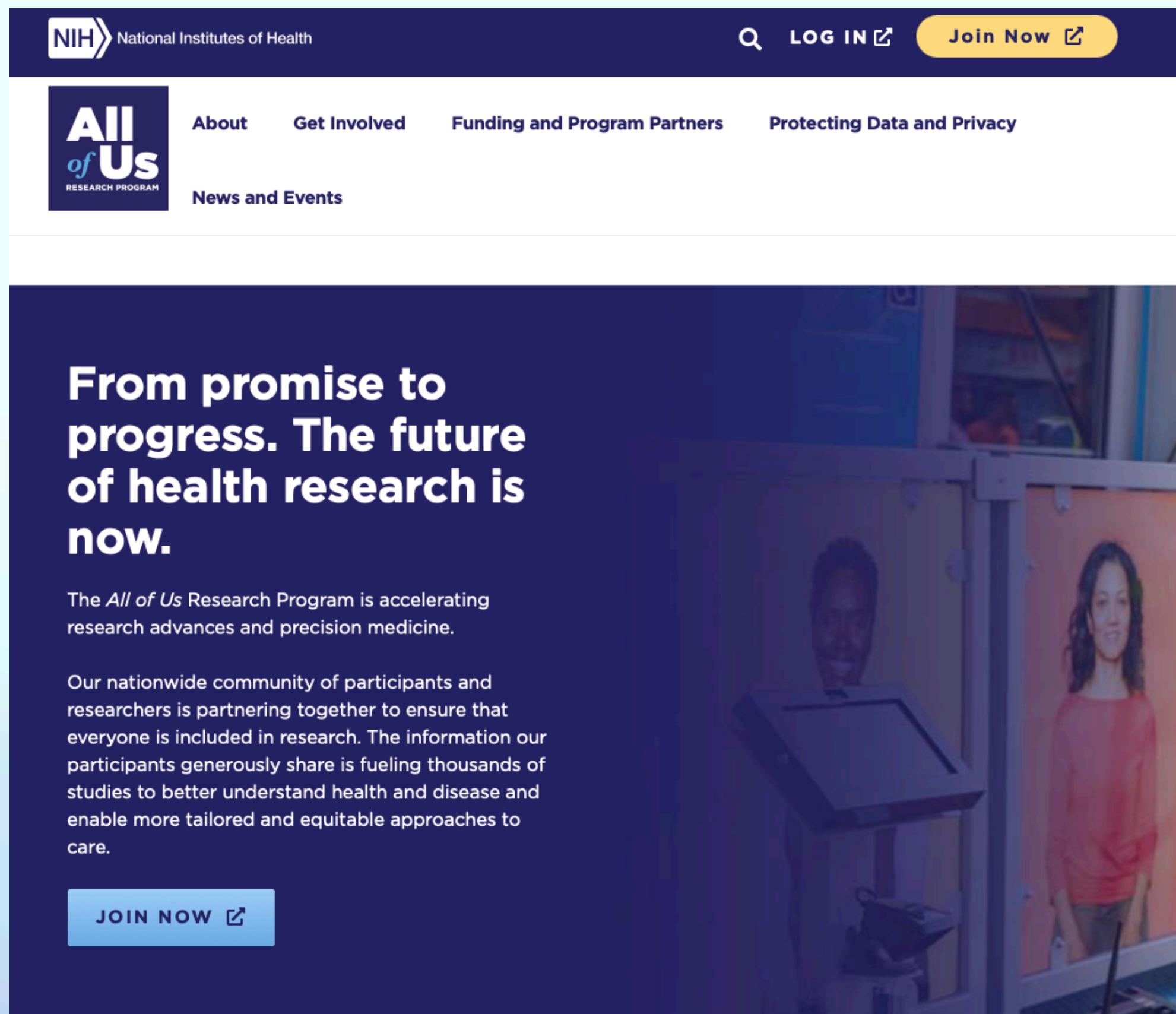
Summary generated 17 May 2024

See under Catalogues for other category groupings.

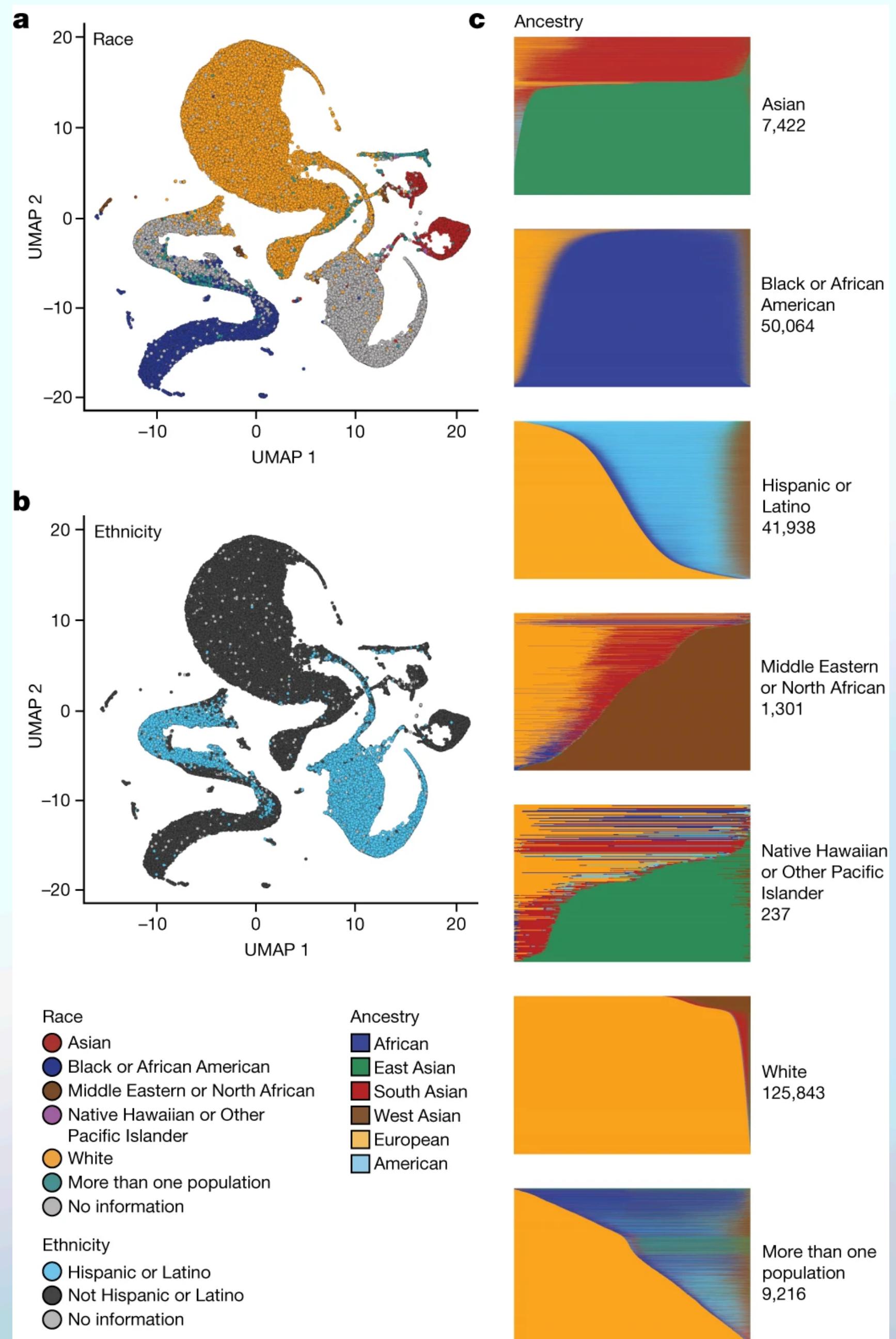
Enabling scientific discoveries that improve human health

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/>

The screenshot shows the HDR-UK website with a navigation bar at the top featuring links to 'Visit the Gateway', 'Visit the Alliance', 'Visit HDR UK Futures', a search bar, and social media icons. Below the navigation is a menu bar with links to 'About', 'Advancing Health Research', 'Access Health Data', 'Helping with Health Data', 'Study and Train', and 'News, Opinion and Events'. The main content area displays a GitHub repository page for 'HDR UK / oss'. The repository details include 5 issues, 1 pull request, 137 commits, 11 branches, 0 tags, and 19 stars. The README file is visible, along with sections for 'Contributors' (susheel, Luis D. Torres) and 'Languages' (Python). At the bottom, there is a section for 'HDR UK Open Source Contributions (173)'.

Our Future Health Genotype Array Data



Metadata last updated: 23/09/2024 [?](#)

Publishing frequency: QUARTERLY [?](#)

[Dataset](#) [Research](#) [United Kingdom](#) [Cohort study](#) [Self-reported health](#) [Show ▾](#)

Our Future Health is a prospective, observational cohort study of the general adult population of the United Kingdom. The dataset includes genotype array data on approximately 700,000 variants from 330,058 participants.

Viewed 423 times

Typical time to access: 1-2 months

Our Future Health Linked Health Records Data



Metadata last updated: 23/09/2024 [?](#)

Publishing frequency: QUARTERLY [?](#)

[Dataset](#) [Research](#) [United Kingdom](#) [Cohort study](#) [Self reported health](#) [Show ▾](#)

Our Future Health is a prospective, observational cohort study of the general adult population of the United Kingdom. This dataset includes linked health records from the NHS in England from 957,444 participants.

Viewed 106 times

Typical time to access: 1-2 months

Our Future Health Baseline Health Questionnaire Data



Metadata last updated: 23/09/2024 [?](#)

Publishing frequency: QUARTERLY [?](#)

[Dataset](#) [Research](#) [Cohort study](#) [United Kingdom](#) [Self-reported health](#) [Show ▾](#)

Our Future Health is a prospective, observational cohort study of the general adult population of the United Kingdom. The dataset includes personal, health and lifestyle information from 1,193,001 participants from a self-completed questionnaire.

Viewed 1086 times

Typical time to access: 1-2 months

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 their specific needs.

NCBI FTP Bulk Download Example

```
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)
```

```
≡ README_ensembl
1  #tax_id ncbi_release  ncbi_assembly  ensembl_release ensembl_assembly
2  3486   Humulus lupulus Annotation Release 100  drHumLupu1.1   Rapid 2024-01-01
3  3517   Alnus glutinosa Annotation Release 100  dhAlnGlut1.1   Rapid 2024-01-01
4  3691   Populus nigra Annotation Release 100   ddPopNigr1.1   Rapid 2024-01-01
5  3986   Mercurialis annua Annotation Release 101   ddMerAnnu1.2   Rapid 2024-01-01
6  6063   Halichondria panicea Annotation Release 100  odHalPani1.1   Rapid 2024-01-01
7  6087   Hydra vulgaris Annotation Release 103   Hydra_105_v3   Rapid 2022-01-01
8  6105   Actinia tenebrosa Annotation Release 100   ASM960242v1 Rapid 2022-01-01
9  6148   0          Rapid 2023-10   ASM1229514v1
10 6454   Haliotis rufescens Annotation Release 101   xgHalRufe1.0.p Rapid 2024-01-01
11 6465   Patella vulgata Annotation Release 100  xgPatVulg1.1   Rapid 2022-01-01
12 6500   Aplysia californica Annotation Release 102   AplCal3.0   Rapid 2021-01-01
13 6526   Biomphalaria glabrata Annotation Release 101   xgBioGlab47.1 Rapid 2024-01-01
14 6549   Mytilus californianus Annotation Release 100   xbMytCal1.0.p Rapid 2024-01-01
15 6565   Crassostrea virginica Annotation Release 100   C_virginica-3.0 Rapid 2024-01-01
16 6573   Mizuhopecten yessoensis Annotation Release 100   ASM211388v2 Rapid 2024-01-01
17 6579   Pecten maximus Annotation Release 100   xPecMax1.1   Rapid 2023-08-01
18 6596   Mercenaria mercenaria Annotation Release 100   ASM1480567v1.1 Rapid 2024-01-01
19 6604   Mya arenaria Annotation Release 100 ASM2691426v1   Rapid 2023-04-01
20 6669   Daphnia pulex Annotation Release 100   ASM2113471v1   Rapid 2022-01-01
21 6687   Penaeus monodon Annotation Release 100  NSTDA_Pmon_1   Rapid NSTDA
22 6689   Penaeus vannamei Annotation Release 100  ASM378908v1 Rapid 2022-04-01
```

KEGG API Download Example

```
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)

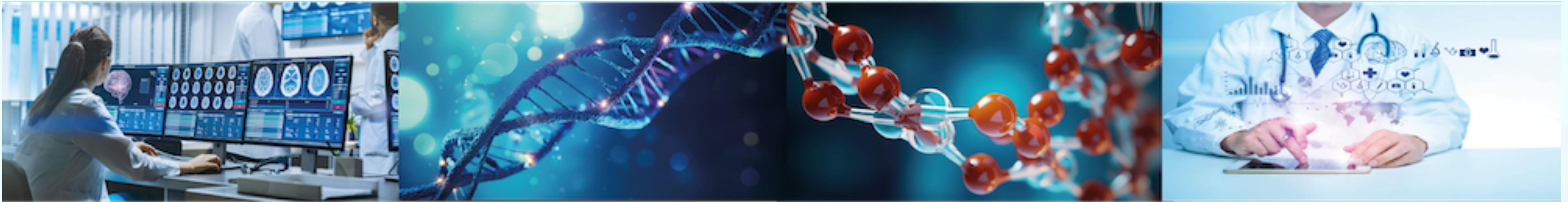
    # Check if the request was successful
    if response.status_code == 200:
        # Write the data content to a file
        with open(f"{pathway_id}.txt", 'w') as file:
            file.write(response.text)
        print(f"Pathway data saved as {pathway_id}.txt")
    else:
        print("Failed to retrieve 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		
1	ENTRY	map00010 Pathway
2	NAME	Glycolysis / Gluconeogenesis
3	DESCRIPTION	Glycolysis is the process of converting glucose into pyruvate and vice versa.
4	CLASS	Metabolism; Carbohydrate metabolism
5	PATHWAY_MAP	map00010 Glycolysis / Gluconeogenesis
6	MODULE	M00001 Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate
7		M00002 Glycolysis, core module involving three-carbon compound
8		M00003 Gluconeogenesis, oxaloacetate => fructose-6P [PATH:map00010]
9		M00307 Pyruvate oxidation, pyruvate => acetyl-CoA [PATH:map00010]
10	DBLINKS	GO: 0006096 0006094
11	REFERENCE	
12	AUTHORS	Nishizuka Y (ed).
13	TITLE	[Metabolic Maps] (In Japanese)
14	JOURNAL	Tokyo Kagaku Dojin (1980)
15	REFERENCE	
16	AUTHORS	Nishizuka Y, Seyama Y, Ikai A, Ishimura Y, Kawaguchi A (eds).
17	TITLE	[Cellular Functions and Metabolic Maps] (In Japanese)
18	JOURNAL	Tokyo Kagaku Dojin (1997)
19	REFERENCE	
20	AUTHORS	Michal G.
21	TITLE	Biochemical Pathways
22	JOURNAL	Wiley (1999)
23	REL_PATHWAY	map00020 Citrate cycle (TCA cycle)
24		map00030 Pentose phosphate pathway
25		map00500 Starch and sucrose metabolism
26		map00620 Pyruvate metabolism
27		map00640 Propanoate metabolism
28		map00710 Carbon fixation by Calvin cycle
29	KO_PATHWAY	k000010
30	///	
31		



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>