JUMPING INTO the Data Curation Stream

Patricia Flatley Brennan, RN, PhD
Director, US National Library of Medicine

Profiles in Science

Nucleotide BLAST
nucleotide ▶ nucleotide

PubMed

MeSH on Demand

WISER

MTI NLM Medical Text Indexer
Biomedical Literature → MTI MTIFL → MeSH Suggestions

ClinicalTrials.gov
A service of the U.S. National Institutes of Health

THE VISIBLE HUMAN PROJECT™
FROM THE NATIONAL LIBRARY OF MEDICINE

NCBI

MEDLINE

NIH NLM › NNLM

CHEMM

TOXNET TOXICOLOGY DATA NETWORK

Genetics Home Reference

LHNCBC

PubMed labs

PubChem

itk

PubMed

Haz-Map @NIH

ACCESS GUDID

LOINC

DAILYMED

VSAC value sets

PubMed Central

OPENi

NLM APIs

GTR Genetic Testing Registry

ClinVar Clinically relevant variation

MedGen Conditions with a genetic component

TOXMAP classic

HISTORY OF MEDICINE

MedlinePlus
Trusted Health Information for You

DOCLINE®

db GaP GENOTYPES and PHENOTYPES

Protein BLAST
protein ▶ protein

U.S. National Library of Medicine

Profiles in Science

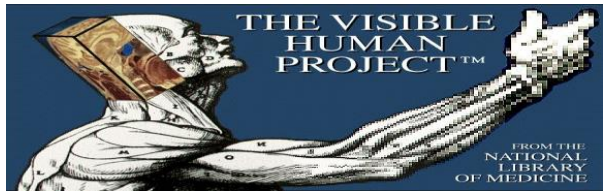
PubMed

MeSH on Demand

WISER

MTI NLM Medical Text Indexer
Biomedical Literature
MTI MTIFL
MeSH Suggestions

ClinicalTrials.gov
A service of the U.S. National Institutes of Health



NCBI

MEDLINE

NIH NLM NNLM

CHEMM

TOXNET TOXICOLOGY DATA NETWORK

Genetics Home Reference

PubChem

itk

PubMed

Haz-Map @NIH

ACCESS GUDID

LHNCBC

PubMed labs

LOINC

DAILYMED

VSAC value sets

PubMed Central

OPENi

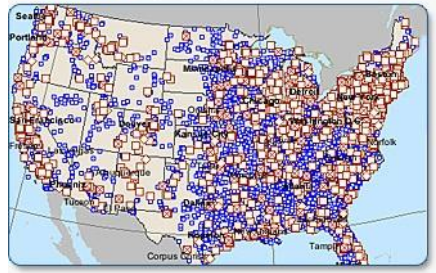
NLM APIs

GTR Genetic Testing Registry

ClinVar Clinically relevant variation

MedGen Conditions with a genetic component

TOXMAP classic



HISTORY OF MEDICINE

MedlinePlus
Trusted Health Information for You

DOCLINE®

db GaP GENOTYPES and PHENOTYPES

Protein BLAST
protein ▶ protein

U.S. National Library of Medicine

Profiles in Science

Nucleotide BLAST
nucleotide ▶ nucleotide

PubMed

MeSH on Demand

WISER

MTI NLM Medical Text Indexer
Biomedical Literature → MTI MTIFL → MeSH Suggestions

ClinicalTrials.gov
A service of the U.S. National Institutes of Health

THE VISIBLE HUMAN PROJECT™
FROM THE NATIONAL LIBRARY OF MEDICINE

NCBI

MEDLINE

NIH NLM › NNLM

CHEMM

TOXNET TOXICOLOGY DATA NETWORK

Genetics Home Reference

LHNCBC

PubMed labs

PubChem

itk

PubMed

Haz-Map @NIH

ACCESS GUDID

LOINC

DAILYMED

VSAC value sets

PubMed Central

OPENi

NLM APIs

GTR Genetic Testing Registry

ClinVar Clinically relevant variation

MedGen Conditions with a genetic component

TOXMAP classic

HISTORY OF MEDICINE

MedlinePlus
Trusted Health Information for You

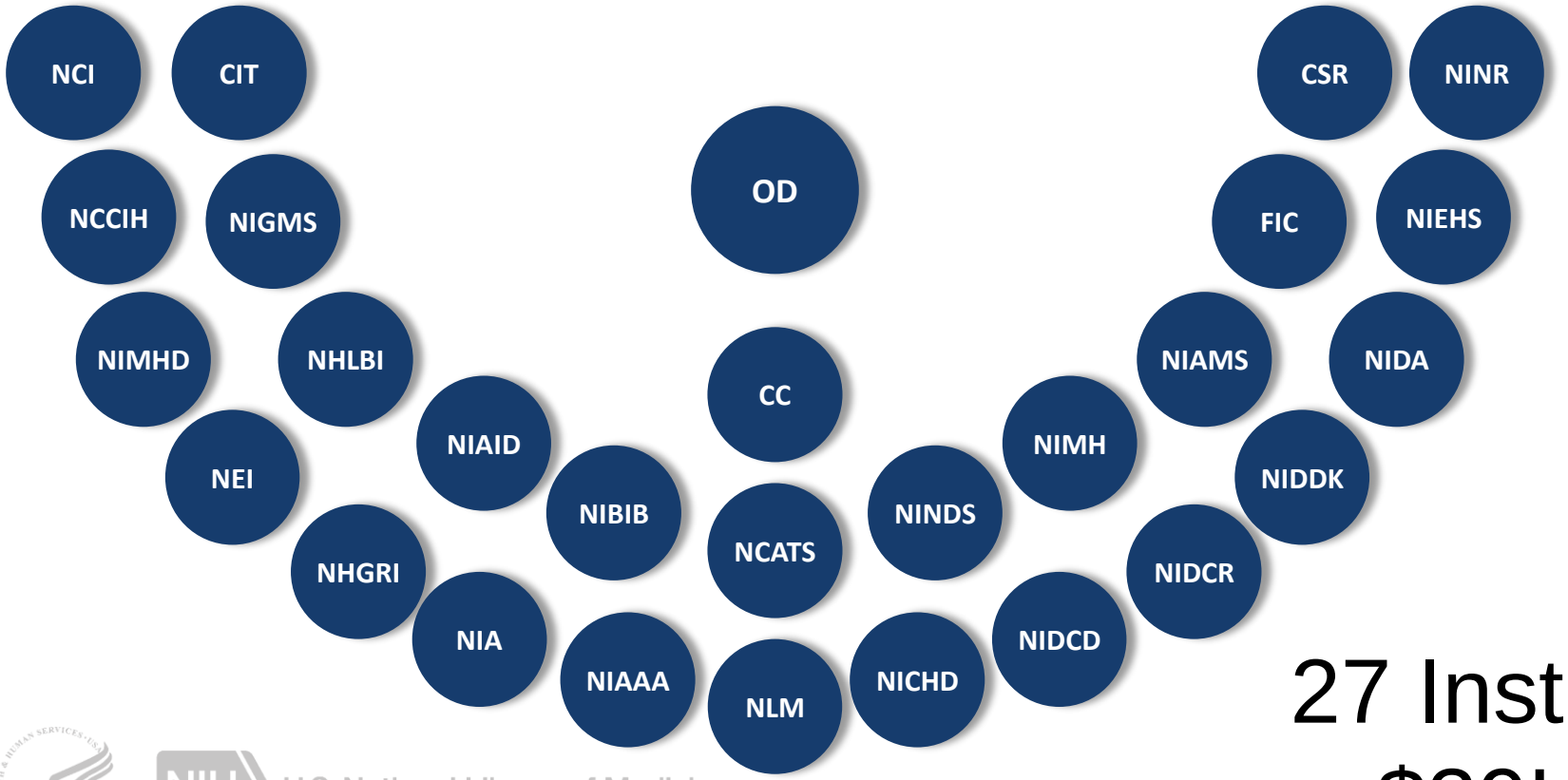
DOCLINE®

db GaP GENOTYPES and PHENOTYPES

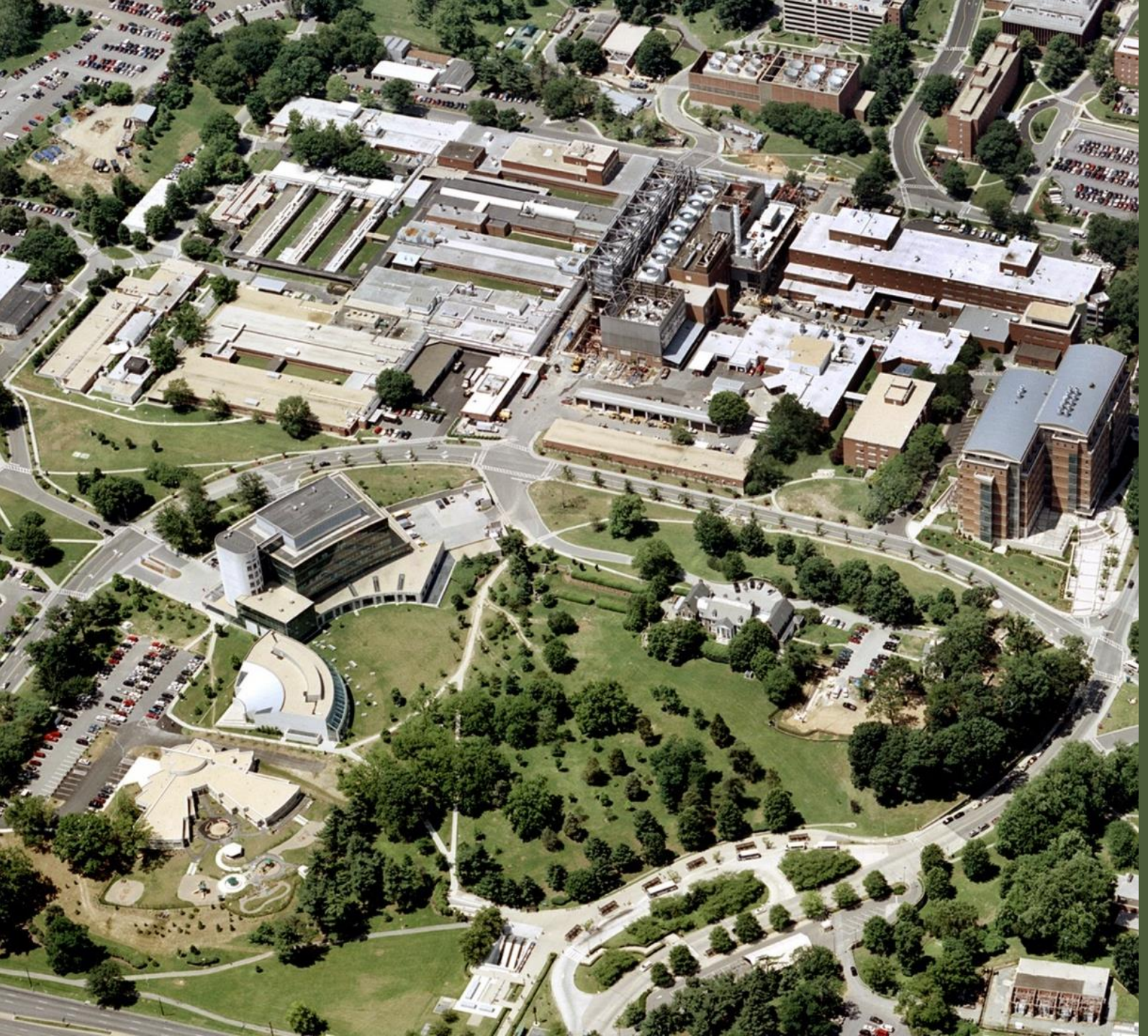
Protein BLAST
protein ▶ protein

U.S. National Library of Medicine

National Institutes of Health



27 Institutes & Centers
$39US billion annually



BIOMEDICAL CHALLENGES

-Cardiovascular Health

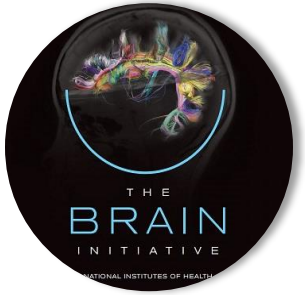
- Gene Therapy

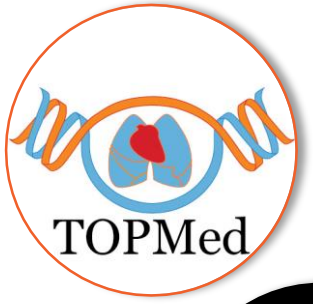
- Alzheimer's Disease

- Lifespan Development

- BRAIN

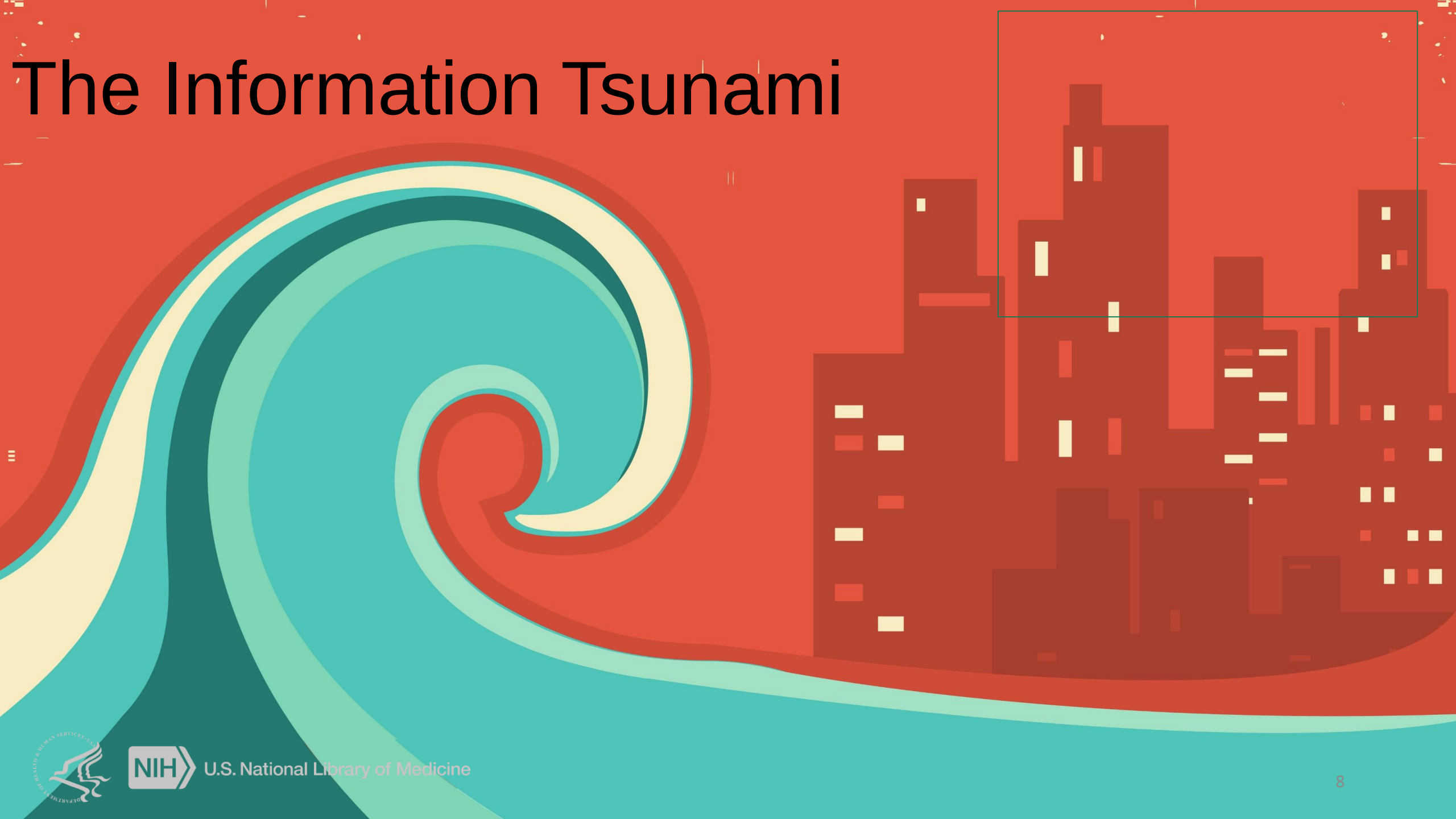
Open Science
Open Data at NIH

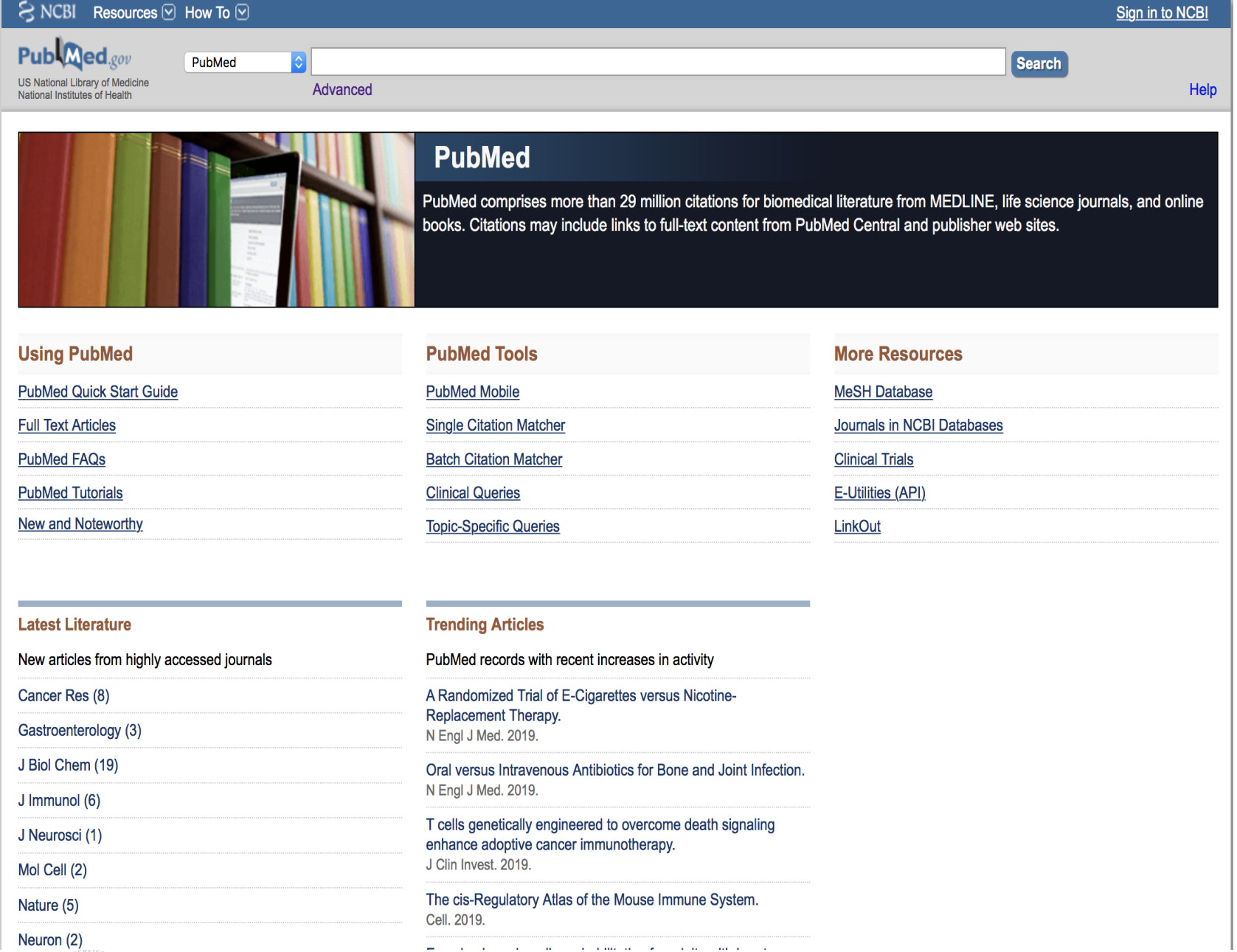




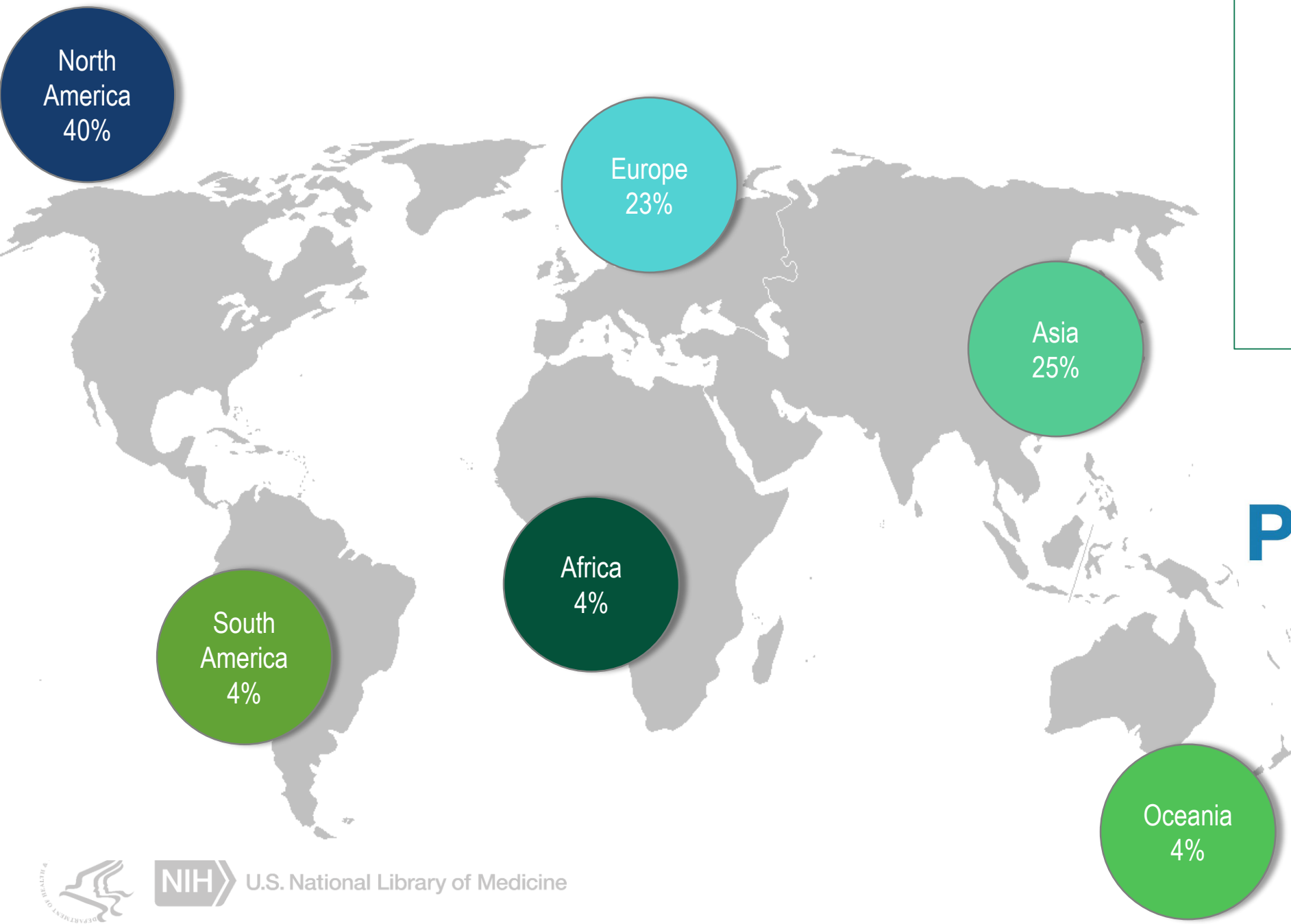
and many others

The Information Tsunami




NCBI Resources How To
Sign in to NCBI
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Search
Advanced
Help
PubMed
PubMed comprises more than 29 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.
Using PubMed
PubMed Quick Start Guide
Full Text Articles
PubMed FAQs
PubMed Tutorials
New and Noteworthy
PubMed Tools
PubMed Mobile
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Topic-Specific Queries
More Resources
MeSH Database
Journals in NCBI Databases
Clinical Trials
E-Utilities (API)
LinkOut
Latest Literature
New articles from highly accessed journals
Cancer Res (8)
Gastroenterology (3)
J Biol Chem (19)
J Immunol (6)
J Neurosci (1)
Mol Cell (2)
Nature (5)
Neuron (2)
Trending Articles
PubMed records with recent increases in activity
A Randomized Trial of E-Cigarettes versus Nicotine-Replacement Therapy.
N Engl J Med. 2019.
Oral versus Intravenous Antibiotics for Bone and Joint Infection.
N Engl J Med. 2019.
T cells genetically engineered to overcome death signaling enhance adoptive cancer immunotherapy.
J Clin Invest. 2019.
The cis-Regulatory Atlas of the Mouse Immune System.
Cell. 2019.

PubMed



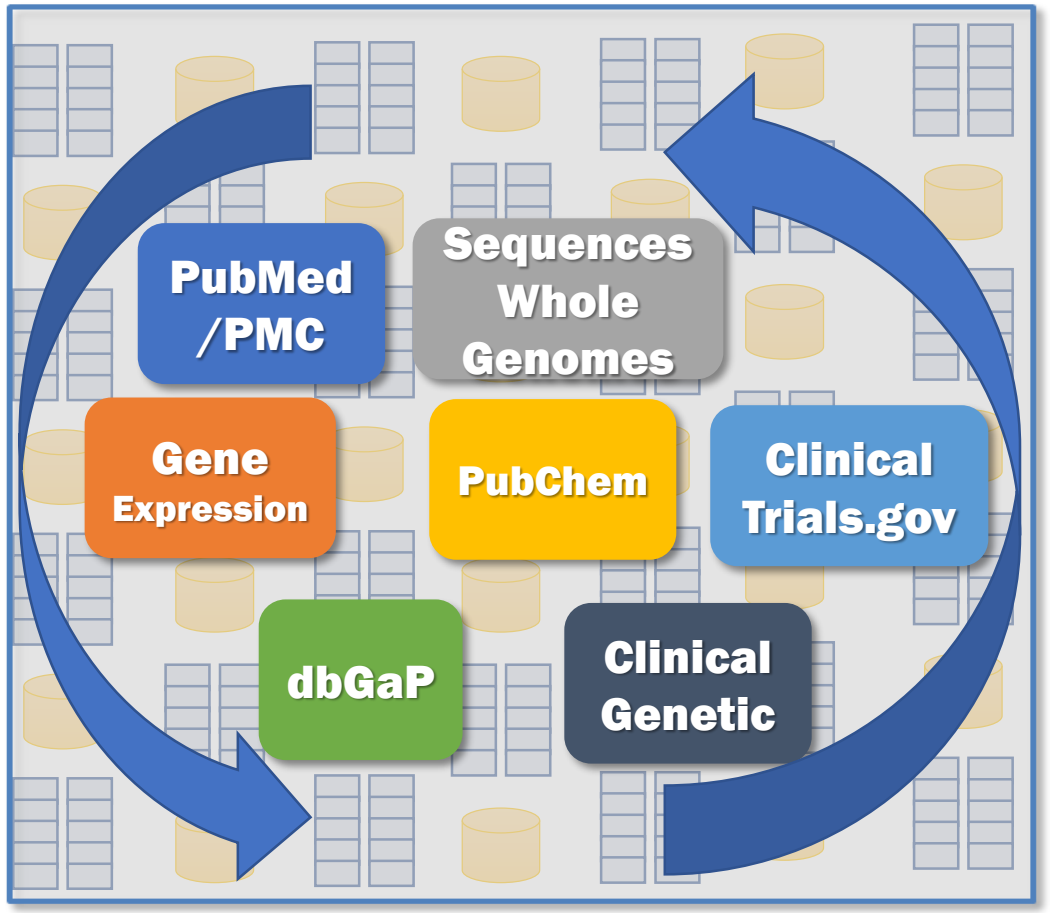
PubMed

Daily Usage

29+ million articles
2.5 million users
~ 3 million searches
9 million page views

Data Resources at NCBI

550 Terabytes/day **200 Petabytes/year**



10-15 Terabytes (TB)day

Data Submissions

20 TB/day

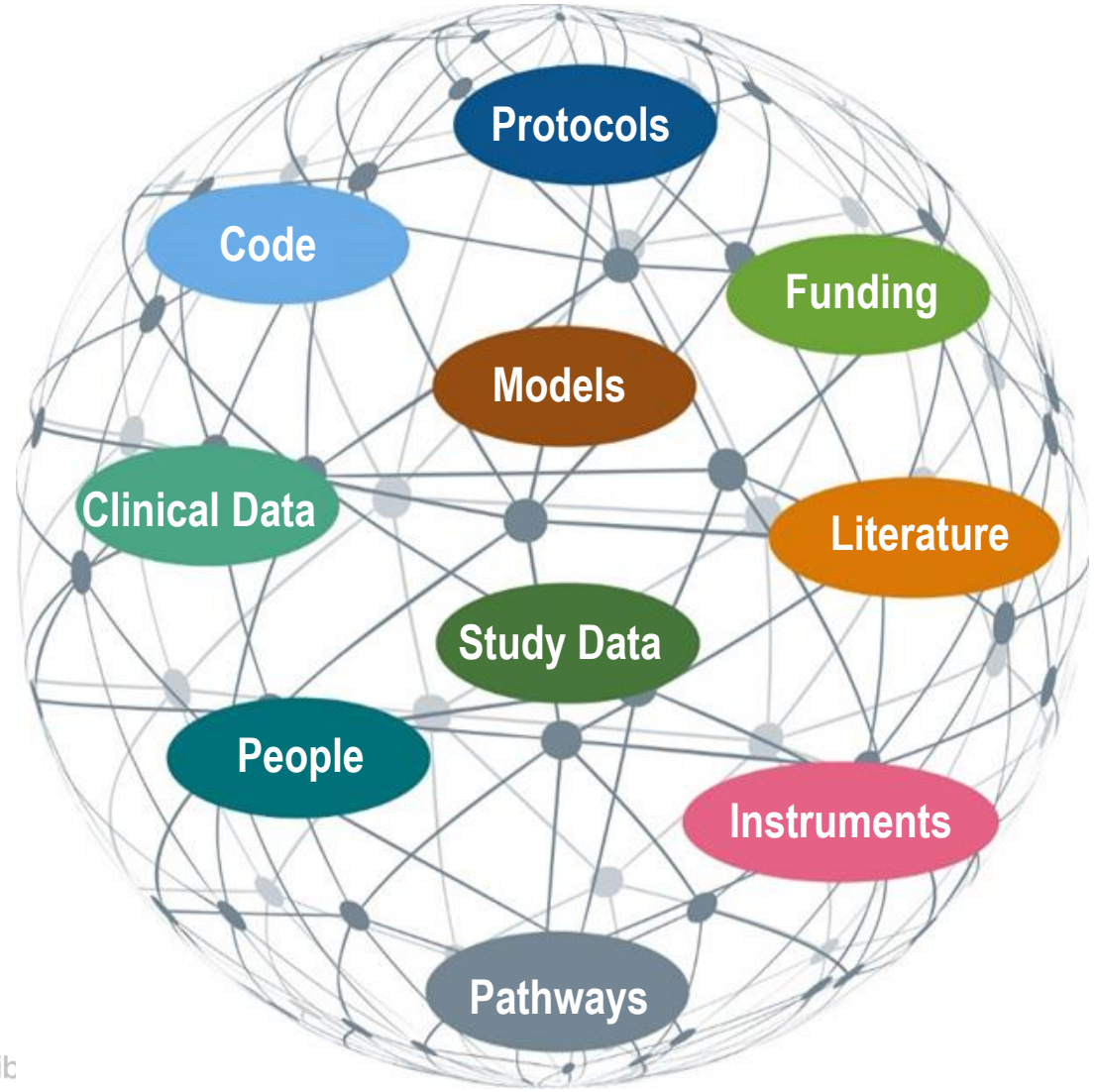
Interactive Web

95 TB/day

Data Downloads

1 Petabyte = 1,000 TB = 1,000,000 GB = 1,500,000 CD-ROMs

Digital Objects Lifecycle



National Library of Medicine

THE

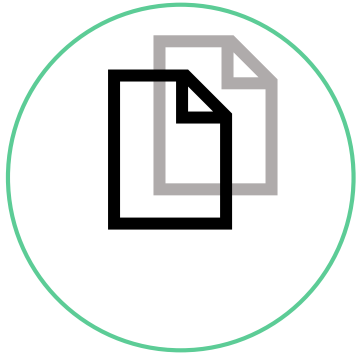
Engine of Curation

Why curate data?

Maximize reuse

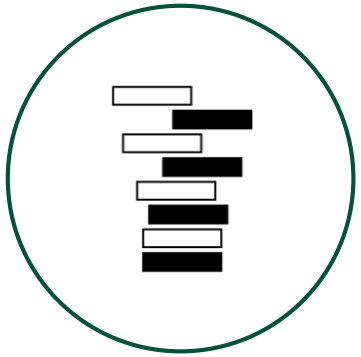
Mitigate obsolescence

Maintain value



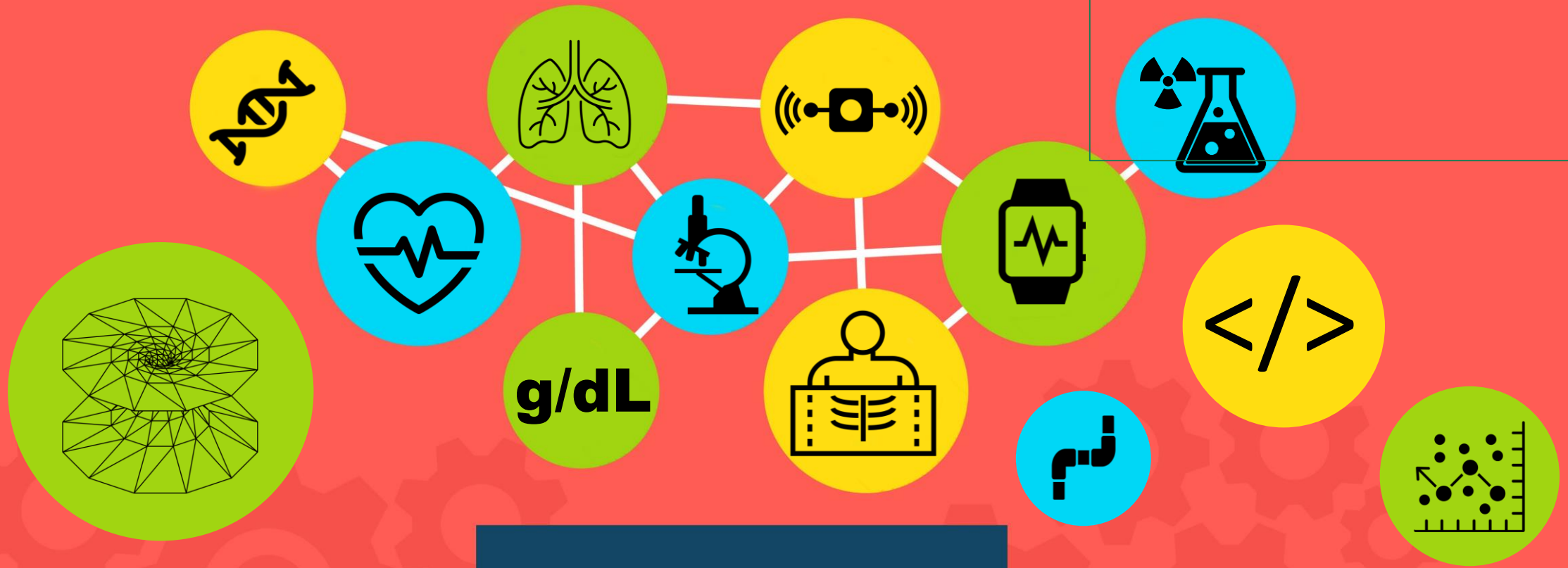
Facilitate reproducibility

Explicate gene-phenotype relations

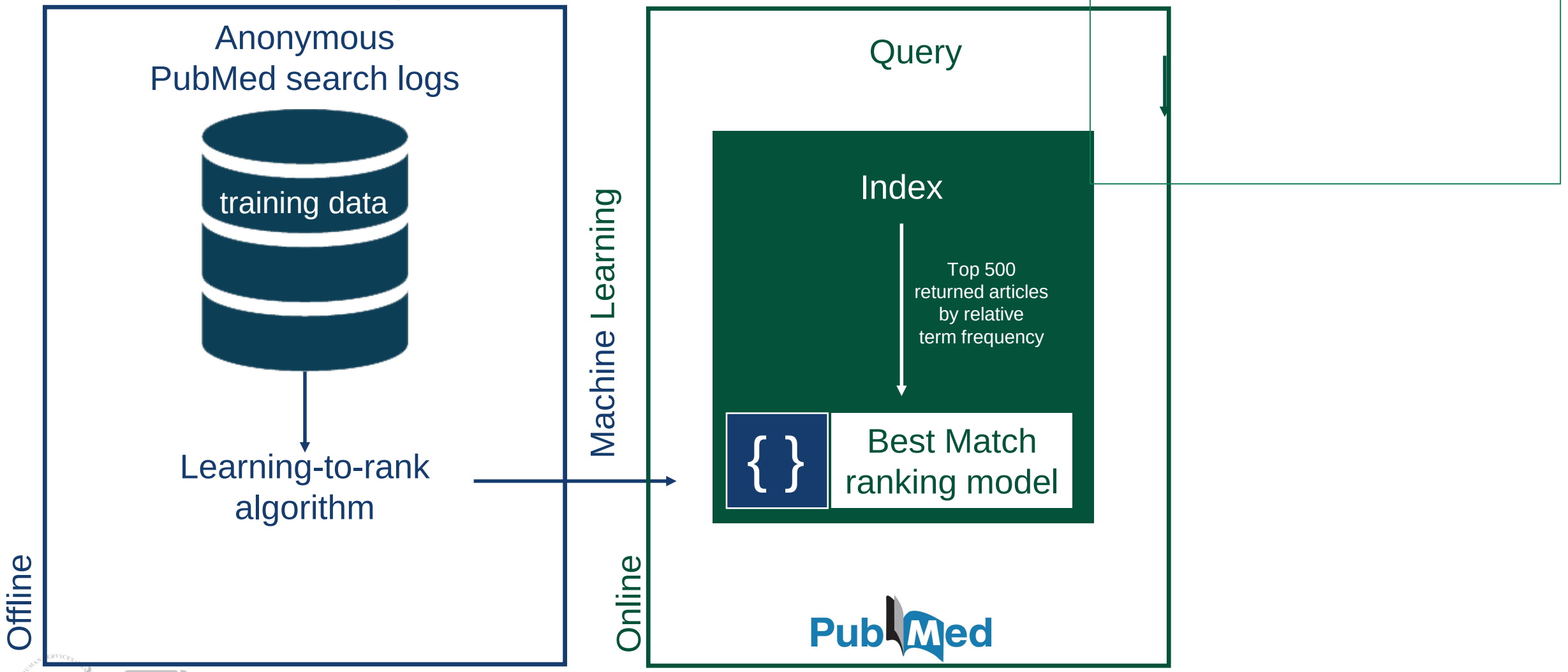


Integrate across repositories

What gets curated?



Leveraging Curation for Discovery



Leveraging Curation for Discovery

Who curates at NLM?

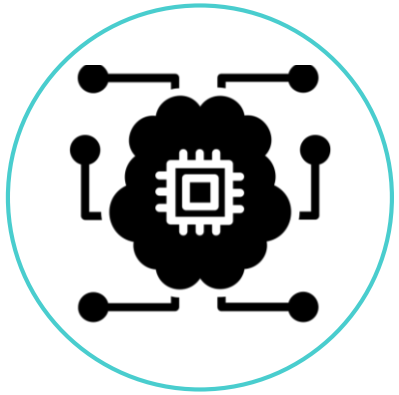
NLM staff

How does curation happen?

< actors

Human effort

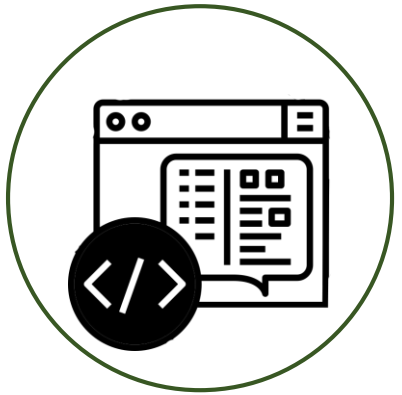
Machine-aided



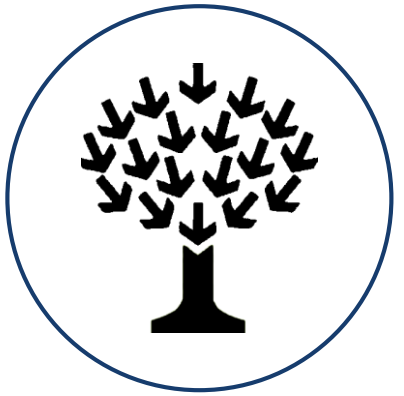
Machine only

tools >

Representation



Encoding

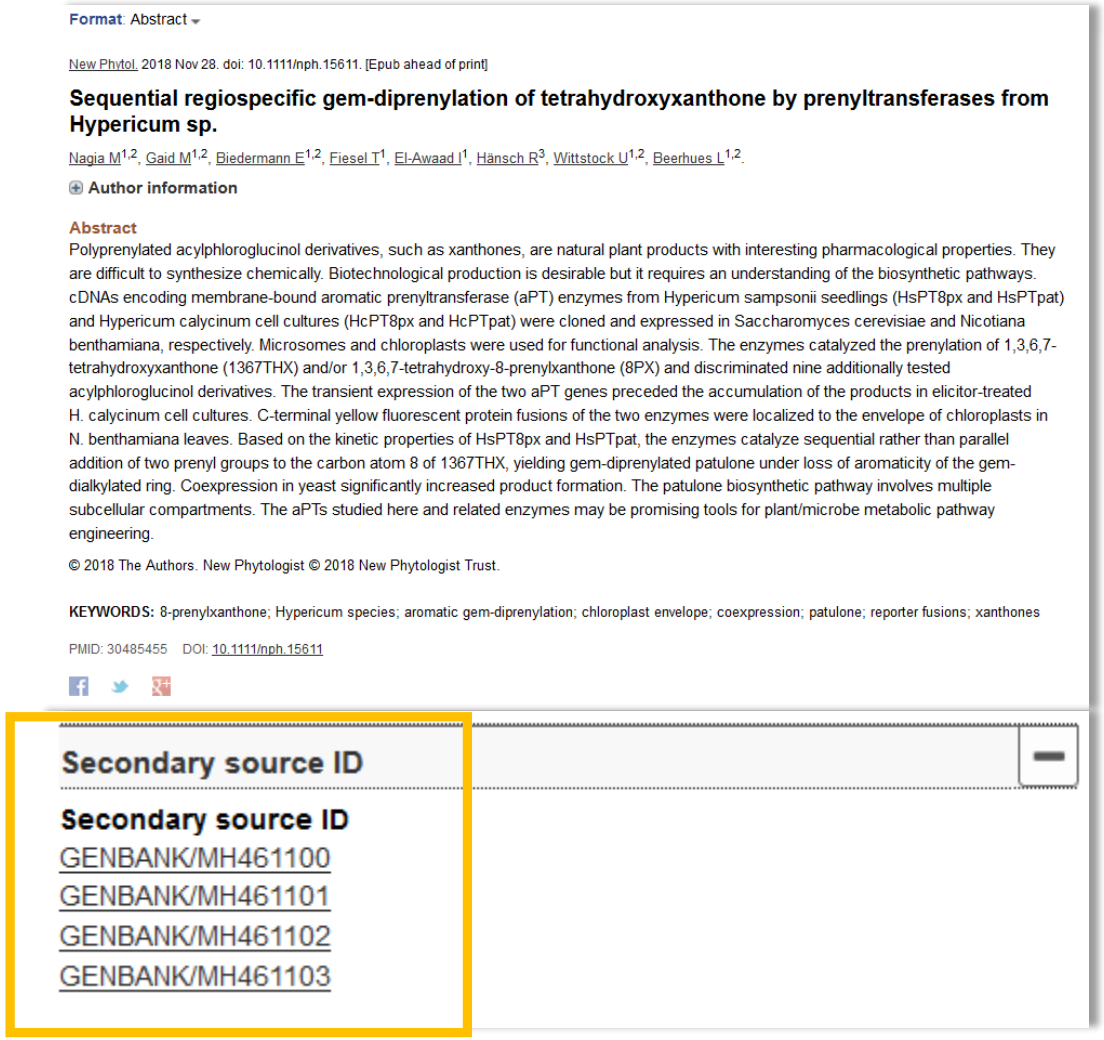


Metathesauri

Data Sharing

in PubMed & PubMed Central

PubMed *Secondary Source*



Format: Abstract

New Phytol. 2018 Nov 28. doi: 10.1111/nph.15611. [Epub ahead of print]

Sequential regiospecific gem-diprenylation of tetrahydroxyxanthone by prenyltransferases from Hypericum sp.

Nagia M[1,2], Gaid M[1,2], Biedermann E[1,2], Fiesel T[1], El-Awaad I[1], Hänsch R[3], Wittstock U[1,2], Beerhues L[1,2].

Author information

Abstract

Polyprenylated acylphloroglucinol derivatives, such as xanthones, are natural plant products with interesting pharmacological properties. They are difficult to synthesize chemically. Biotechnological production is desirable but it requires an understanding of the biosynthetic pathways. cDNAs encoding membrane-bound aromatic prenyltransferase (aPT) enzymes from Hypericum sampsonii seedlings (HsPT8px and HsPTpat) and Hypericum calycinum cell cultures (HcPT8px and HcPTpat) were cloned and expressed in Saccharomyces cerevisiae and Nicotiana benthamiana, respectively. Microsomes and chloroplasts were used for functional analysis. The enzymes catalyzed the prenylation of 1,3,6,7-tetrahydroxyxanthone (1367THX) and/or 1,3,6,7-tetrahydroxy-8-prenylxanthone (8PX) and discriminated nine additionally tested acylphloroglucinol derivatives. The transient expression of the two aPT genes preceded the accumulation of the products in elicitor-treated H. calycinum cell cultures. C-terminal yellow fluorescent protein fusions of the two enzymes were localized to the envelope of chloroplasts in N. benthamiana leaves. Based on the kinetic properties of HsPT8px and HsPTpat, the enzymes catalyze sequential rather than parallel addition of two prenyl groups to the carbon atom 8 of 1367THX, yielding gem-diprenylated patulone under loss of aromaticity of the gem-dialkylated ring. Coexpression in yeast significantly increased product formation. The patulone biosynthetic pathway involves multiple subcellular compartments. The aPTs studied here and related enzymes may be promising tools for plant/microbe metabolic pathway engineering.

KEYWORDS: 8-prenylxanthone; Hypericum species; aromatic gem-diprenylation; chloroplast envelope; coexpression; patulone; reporter fusions; xanthones

PMID: 30485455 DOI: 10.1111/nph.15611

Secondary source ID

Secondary source ID

GENBANK/MH461100
GENBANK/MH461101
GENBANK/MH461102
GENBANK/MH461103

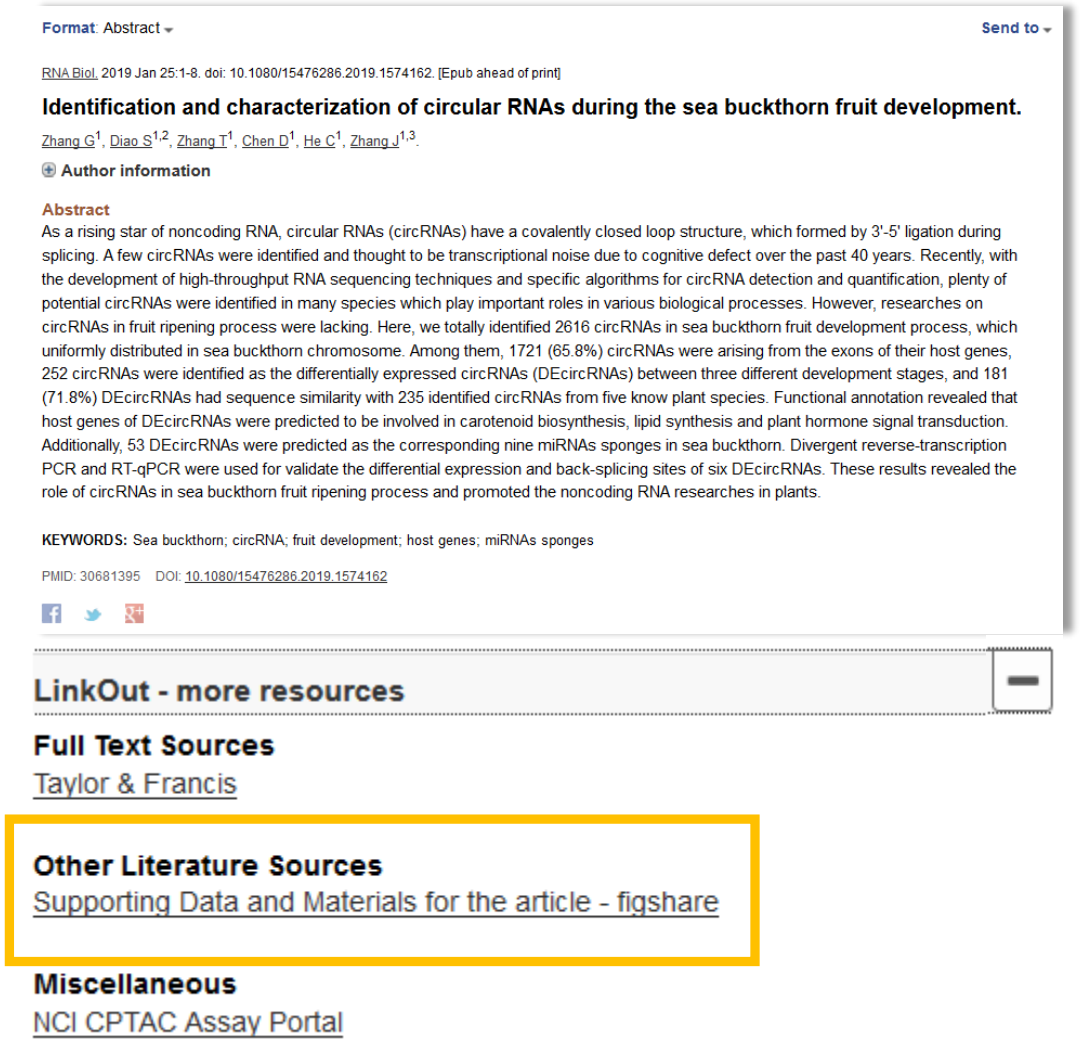
Data Sources

- Publishers
- NLM indexers
- PubMed Central (PMC)

Search

- has suppdata[filter]
- genbank[si]
- figshare[si]

PubMed *Linked Data*



Format: Abstract — Send to

RNA Biol. 2019 Jan 25:1-8. doi: 10.1080/15476286.2019.1574162. [Epub ahead of print]

Identification and characterization of circular RNAs during the sea buckthorn fruit development.

Zhang G[1], Diao S[1,2], Zhang T[1], Chen D[1], He C[1], Zhang J[1,3].

Author information

Abstract

As a rising star of noncoding RNA, circular RNAs (circRNAs) have a covalently closed loop structure, which formed by 3'-5' ligation during splicing. A few circRNAs were identified and thought to be transcriptional noise due to cognitive defect over the past 40 years. Recently, with the development of high-throughput RNA sequencing techniques and specific algorithms for circRNA detection and quantification, plenty of potential circRNAs were identified in many species which play important roles in various biological processes. However, researches on circRNAs in fruit ripening process were lacking. Here, we totally identified 2616 circRNAs in sea buckthorn fruit development process, which uniformly distributed in sea buckthorn chromosome. Among them, 1721 (65.8%) circRNAs were arising from the exons of their host genes, 252 circRNAs were identified as the differentially expressed circRNAs (DEcircRNAs) between three different development stages, and 181 (71.8%) DEcircRNAs had sequence similarity with 235 identified circRNAs from five know plant species. Functional annotation revealed that host genes of DEcircRNAs were predicted to be involved in carotenoid biosynthesis, lipid synthesis and plant hormone signal transduction. Additionally, 53 DEcircRNAs were predicted as the corresponding nine miRNAs sponges in sea buckthorn. Divergent reverse-transcription PCR and RT-qPCR were used for validate the differential expression and back-splicing sites of six DEcircRNAs. These results revealed the role of circRNAs in sea buckthorn fruit ripening process and promoted the noncoding RNA researches in plants.

KEYWORDS: Sea buckthorn; circRNA; fruit development; host genes; miRNAs sponges

PMID: 30681395 DOI: 10.1080/15476286.2019.1574162

LinkOut - more resources

Full Text Sources

Taylor & Francis

Other Literature Sources

Supporting Data and Materials for the article - figshare

Miscellaneous

NCI CPTAC Assay Portal

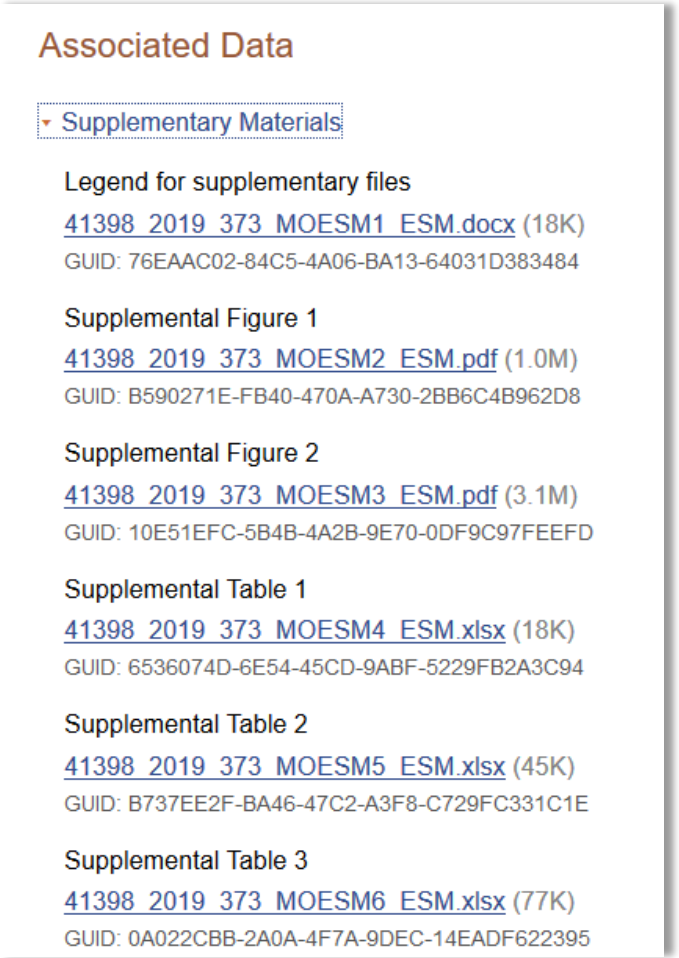
Supporting Materials

- data sets
- graphics
- images
- sound
- multimedia

Search

- loprovfigshare[filter]
- loprovdryaddb[filter]

PubMed Central *Supplementary Data*



What you might find

- Computer code
- Mathematical or computational models
- Audio or video clips
- Oversized tables
- Intervention protocols
- Primary or supplementary data sets
- Expanded methodology sections
- Additional figures

Source: Publication Manual of APA (6th ed.)

Search: has suppdata[filter]

Creating a Library of Models

Why care about models?

igor

Characterizing Models

- Data elements
  - Model type
  - Purpose
  - Assumptions
  - Use
  - Scale
- Model verification
- Model validation

Starting Point: Unique IDs

- Machine readable
- Free to generate & access
- Persistent
- Consistent across journals & publishers
- Public registration authorities
  - RRID: scicrunch.org
  - DOI: doi.org

RRID and DOI's

RRIDs > > Methods

- model organisms
- cell lines
- antibodies
- software
- databases
- algorithms
- blogs

DOIs >> Objects

- articles
- books
- datasets
- objects

PubMed

J Comp Neurol. 2016 Apr 15;524(6):1236-58. doi: 10.1002/cne.23901. Epub 2015 Sep 29.

DOI

Brain region-dependent differential expression of alpha-synuclein.

Taguchi K[1], Watanabe Y[1], Tsujimura A[1], Tanaka M[1].

Author information

Abstract

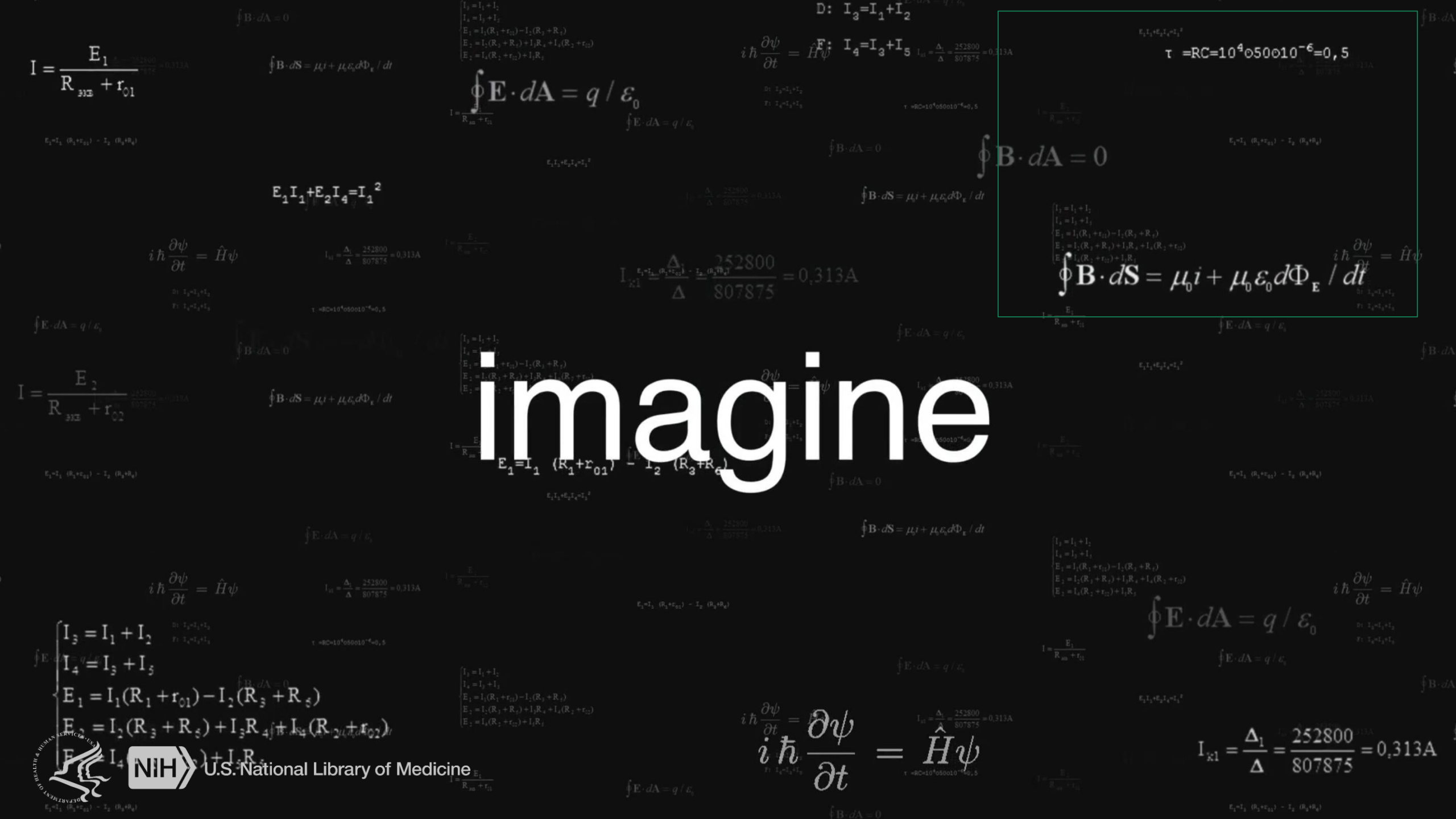
α-Synuclein, the major constituent of Lewy bodies (LBs), is normally expressed in presynapses and is involved in synaptic function. Abnormal intracellular aggregation of α-synuclein is observed as LBs and Lewy neurites in neurodegenerative disorders, such as Parkinson's disease (PD) or dementia with Lewy bodies. Accumulated evidence suggests that abundant intracellular expression of α-synuclein is one of the risk factors for pathological aggregation. Recently, we reported differential expression patterns of α-synuclein between excitatory and inhibitory hippocampal neurons. Here we further investigated the precise expression profile in the adult mouse brain with special reference to vulnerable regions along the progression of idiopathic PD. The results show that α-synuclein was highly expressed in the neuronal cell bodies of some early PD-affected brain regions, such as the olfactory bulb, dorsal motor nucleus of the vagus, and substantia nigra pars compacta. Synaptic expression of α-synuclein was mostly accompanied by expression of vesicular glutamate transporter-1, an excitatory presynaptic marker. In contrast, expression of α-synuclein in the GABAergic inhibitory synapses was different among brain regions. α-Synuclein was clearly expressed in inhibitory synapses in the external plexiform layer of the olfactory bulb, globus pallidus, and substantia nigra pars reticulata, but not in the cerebral cortex, subthalamic nucleus, or thalamus. These results suggest that some neurons in early PD-affected human brain regions express high levels of perikaryal α-synuclein, as happens in the mouse brain. Additionally, synaptic profiles expressing α-synuclein are different in various brain regions.

KEYWORDS: GAD; Parkinson's disease; RRID: AB_10048713; RRID: AB_2192953; RRID: AB_2199314; RRID: AB_2301751; RRID: AB_390204; RRID: AB_398108; RRID: AB_477019; mouse; synapse; vGluT-1

PMID: 26358191 DOI: 10.1002/cne.23901

[Indexed for MEDLINE]

RRIDs



Curation AT NLM

Now … into the future

Who curates at NLM?

NLM staff

Data stewards

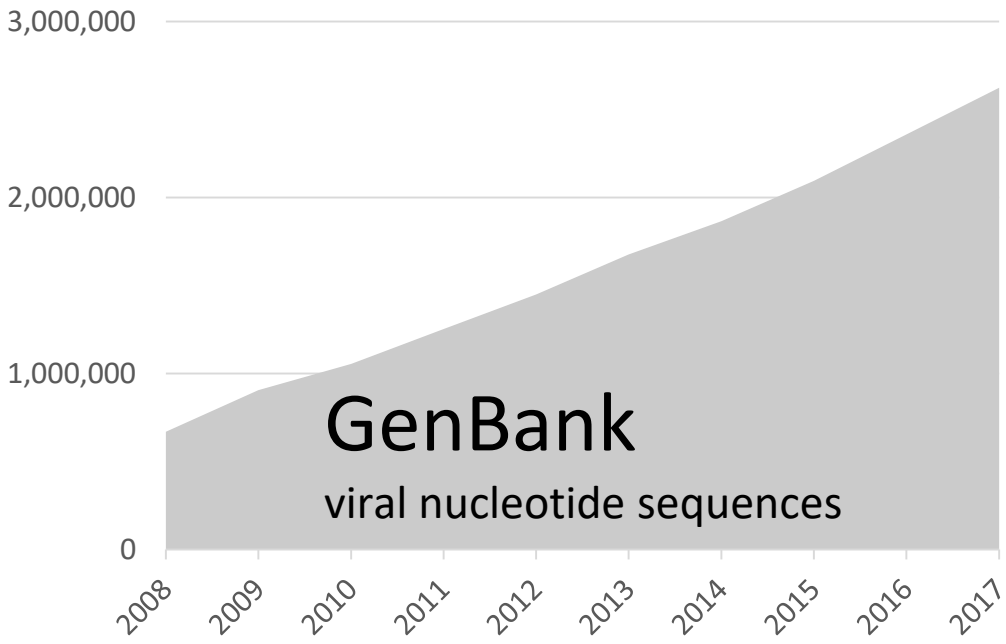
Specialists

Generalists

Partners

GenBank Virus Sequences

The high volume of submissions offers great potential for scientific discoveries & public health interventions.



TATGCTCGCTGATCGAGCGTGATAGATCAGCTATC
GACTAGAGTCACTGACT GACTAGAGTCACTGACT
TATGCTCGCTGATCGAGCGTGATAGATCAGCTATC

Nucleotide sequences need up-to-date, standardized annotations to be useful.

- Validate sequence assembly
- Translate protein sequences
- Detect important mutations

TATGCTCGCTGATCGAGCGTGATAGATCAGCTATC
GACTAGAGTCACTGACT GACTAGAGTCACTGACT

Annotation Tool

for Sequence Submissions

Test case: Influenza viruses

Goals: Better data, happy data submitters

Web Wizard and programmatic interface:

- Validate taxonomy
- Identify sequence errors
- Annotate proteins
- Validate host, strain, and other metadata
- Generate isolate name from metadata

Outcomes: Cut turn-around time from days to minutes

678 submissions processed
44,351 sequences published
since January 2018


NIH
U.S. National Library of Medicine
A Platform for Biomedical Discovery and Data-Powered Health
Strategic Plan 2017-2027
NIAMS
NIAID
NEI
NIDCR
CC
NCI
NLM

Foster the distinctiveness of NLM as a

reliable, trustable source

of health information & biomedical data

@NLMdirector

patti.brennan@nih.gov

nlmdirector.nlm.nih.gov

CREDITS

Artificial Intelligence | ProSymbols
Checkmark | Arif Fajar Yulianto
Coding | Gregor Cresnar
Coding | Iconstock
Coin | Kirill Tomilov
Dinosaur | Rayhan Maulana Rikzan
Due Diligence | Gregor Cresnar
Fingerprint | Elves Sousa
Heartbeat | Deemak Daksina
Heart Monitor Watch| Sergey Demushkin
Inner Join | David Christensen
Integration | Theresa Stoodley

The Noun Project

Laboratory | Alone Forever
Lungs | Mahmure Alp
Microscope | Ricardo Rojas Parametric
Design | Lluisa Iborra
Partner | Lipi
Regression Analysis | Vectors Market
Sensor | Carolina Cani
Specialist | Tuong Tam
Team | Luis Prado
Tuxedo | Micah Lawrence
Water Pipeline | ProSymbols
Work | BGBOXXX Design
X-Ray | Rawisoot Narksrisook