Data From the National Library of Medicine

UNT Open Access Symposium 2017

Daniel E. Burgard
University of North Texas Health Science Center at Fort Worth
Regional Medical Library, South Central Region, National Library of Medicine
May 19, 2017
• “Serving as a leading global resource for building, curating and providing sophisticated access to molecular biology and genomic information, including those from the Human Genome Project and the NIH Common Fund;”

• “Creating high-quality information services relevant to toxicology and environmental health, health services research, and public health;”
NLM Regional Medical Libraries
2016 - 2021

Host Libraries:

Univ of Massachusetts
Univ of Pittsburgh
Univ of Maryland
Univ Iowa
Univ of Utah
UNTHSC
Univ of Washington
UCLA
Data Types
- Bibliographic
- Toxicological
- Chemical
- Health
- Genetic

Audiences
- Basic scientists / researchers
- Healthcare professionals
- General public
- All ages
- Many languages
▪ Data and systems produced by NLM & NIH employees
▪ Data produced by NIH awardees
From Bench to Bedside
PubMed Central® (PMC) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM).

Get Started
- PMC Overview
- Users' Guide
- Journal List
- PMC FAQs
- PMC Copyright Notice

Participate
- Information for Publishers
- Add a Journal to PMC
- Participation Agreements
- File Submission Specifications
- File Validation Tools

Keep Up to Date
- New in PMC | RSS
- PMC Announce Mail List
- Utilities Announce Mail List
- Tagging Guidelines Mail List

Other Resources
- PMC International
- Text Mining Collections
- Developer Resources
- NLM LitArch

Public Access
- Funders and PMC
- How Papers Get Into PMC
- NIH Manuscript Submission System
- My Bibliography

4.3 MILLION Articles are archived in PMC.

Content provided in part by:

2007  326  4247
Genes are the building blocks of inheritance. Passed from parent to child, they contain instructions for making proteins. If genes don't produce the right proteins or don't produce them correctly, a child can have a genetic disorder.
<table>
<thead>
<tr>
<th>Literature</th>
<th>Genes</th>
<th>Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>EST</td>
<td>ClinVar</td>
</tr>
<tr>
<td>MeSH</td>
<td>Gene</td>
<td>dbGaP</td>
</tr>
<tr>
<td>NLM Catalog</td>
<td>GEO DataSets</td>
<td>QTR</td>
</tr>
<tr>
<td>PubMed</td>
<td>GEO Profiles</td>
<td>MedGen</td>
</tr>
<tr>
<td>PubMed Central</td>
<td>HomoLogene</td>
<td>OMIM</td>
</tr>
<tr>
<td></td>
<td>PopSet</td>
<td>PubMed Health</td>
</tr>
<tr>
<td></td>
<td>UniGene</td>
<td>Genomes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assembly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BioProject</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BioSample</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clone</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dbVar</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Genome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nucleotide</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Probe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SNP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SRA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Taxonomy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Literature**: books and reports
- **Books**: ontology used for PubMed indexing
- **MeSH**: books, journals and more in the NLM Collections
- **NLM Catalog**: scientific & medical abstracts/citations
- **PubMed**: full-text journal articles

- **Genes**: expressed sequence tag sequences
- **EST**: collected information about gene loci
- **Gene**: functional genomics studies
- **GEO DataSets**: gene expression and molecular abundance profiles
- **GEO Profiles**: homologous gene sets for selected organisms
- **HomoLogene**: sequence sets from phylogenetic and population studies
- **PopSet**: clusters of expressed transcripts

- **Health**: human variations of clinical significance
- **ClinVar**: genotype-phenotype interaction studies
- **dbGaP**: genetic testing registry
- **QTR**: medical genetics literature and links
- **MedGen**: online mendelian inheritance in man
- **OMIM**: clinical effectiveness, disease and drug reports

- **Proteins**: conserved protein domains
- **Conserved Domains**: protein sequences
- **Protein**: sequence similarity-based protein clusters
- **Protein Clusters**: experimentally-determined biomolecular structures

- **Genomes**: genome assembly information
- **Assembly**: biological projects providing data to NCBI
- **BioProject**: descriptions of biological source materials
- **BioSample**: genomic and cDNA clones
- **Clone**: genome structural variation studies
- **dbVar**: genome sequencing projects by organism
- **Genome**: genome survey sequences
- **GSS**: DNA and RNA sequences
- **Nucleotide**: sequence-based probes and primers
- **Probe**: short genetic variations
- **SNP**: high-throughput DNA and RNA sequence read archive
- **SRA**: taxonomic classification and nomenclature catalog

- **Chemicals**: molecular pathways with links to genes, proteins and chemicals
- **BioSystems**: bioactivity screening studies
- **PubChem BioAssay**: chemical information with structures, information and links
- **PubChem Compound**: deposited substance and chemical information
- **PubChem Substance**:
Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

Web BLAST

Nucleotide BLAST
nucleotide → nucleotide

blastx
translated nucleotide → protein

tblastn
protein → translated nucleotide

Protein BLAST
protein → protein

BLAST Genomes

Enter organism common name, scientific name, or tax ID

Search

Human  Mouse  Rat  Microbes

Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider
Genome data download made easy!

Posted on May 8, 2017

This blog post is directed toward Assembly users.

A new "Download assemblies" button is now available in the Assembly database. This makes it easy to download data for multiple genomes without having to write scripts.

For example, you can run a search in Assembly and use check boxes (see left side of screenshot below) to refine the set of genome assemblies of interest. Then, just open the "Download assemblies" menu, choose the source database (GenBank or RefSeq), choose the file type, and start the download. An archive file will be saved to your computer that can be expanded into a folder containing your selected genome data files.
NIH Data
### NIH Data Sharing Repositories

This table lists NIH-supported data repositories that make data accessible for reuse. Most accept submissions of appropriate data from NIH-funded investigators (and others), but some restrict data submission to only those researchers involved in a specific research network. Also included are resources that aggregate information about biomedical data and information sharing systems. The table can be sorted according by name and by NIH Institute or Center and may be searched using keywords so that you can find repositories more relevant to your data. Links are provided to information about submitting data to and accessing data from the listed repositories. Additional information about the repositories and points-of-contact for further information or inquiries can be found on the websites of the individual repositories. Are we missing a data sharing repository? Contact us.

#### Show 50 entries

<table>
<thead>
<tr>
<th>IC</th>
<th>Repository Name</th>
<th>Repository Description</th>
<th>Data Submission Policy</th>
<th>Access to Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC1</td>
<td>Cancer Nanotechnology Laboratory (caNanoLab)</td>
<td>caNanoLab is a data sharing portal designed to facilitate information sharing in the biomedical nanotechnology research community to expedite and validate the use of nanotechnology in medicine. caNanoLab provides support for the annotation of nanomaterials with characteristics resulting from physico-chemical, in vitro, and in vivo assays and the sharing of these characteristics and associated nanotechnology protocols in a secure fashion.</td>
<td>How to submit your data to caNanoLab</td>
<td>How to access caNanoLab data</td>
</tr>
<tr>
<td>NC1</td>
<td>The Cancer Imaging Archive (TCIA)</td>
<td>The image data in The Cancer Imaging Archive (TCIA) is organized into purpose-built collections of subjects. The subjects typically have a cancer type and/or anatomical site (lung, brain, etc.) in common.</td>
<td>How to submit data to TCIA</td>
<td>How to access TCIA data</td>
</tr>
<tr>
<td>NC1</td>
<td>PeptideAtlas</td>
<td>PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments. Mass spectrometer output files are collected for human, mouse, yeast, and several other organisms, and searched using the latest search engines and protein sequences.</td>
<td>How to submit data to PeptideAtlas</td>
<td>How to access PeptideAtlas data</td>
</tr>
<tr>
<td>NEI</td>
<td>EyeBENER</td>
<td>The eyeBENER Biorepository and corresponding Database contains family history and clinical eye exam data from subjects enrolled in eyeBENER Program coupled to clinical grade DNA samples. This data and samples are submitted by collaborators throughout the US and Canada and the data is available on a controlled access basis to researchers worldwide.</td>
<td>How to submit data to EyeBENER</td>
<td>How to access EyeBENER data</td>
</tr>
<tr>
<td>NHGRI</td>
<td>FlyBase, A Drosophila Genetic and Genetic Database</td>
<td>Drosophila Genomic and Genetic database that includes proteomics data, microarrays and Tiling BACs.</td>
<td>How to submit data to Flybase</td>
<td>How to access Flybase data</td>
</tr>
<tr>
<td>NHGRI</td>
<td>The Zebrafish Model Organism Database (ZFIN)</td>
<td>ZFIN serves as the zebrafish model organism database. It aims to: a) be the community database resource for the laboratory use of zebrafish, b) develop and support integrated zebrafish genetic, genomic and developmental information, c) maintain the definitive reference data sets of zebrafish research information, d) to link this information extensively to corresponding data in other model organism and human databases, e)</td>
<td>How to submit data to ZFIN</td>
<td>How to access ZFIN data</td>
</tr>
</tbody>
</table>
 NIH Data Sharing Policies

This table lists data sharing policies in effect at NIH. It includes policies at the NIH, IC, division, and program levels that apply to broad sets of investigators and data. Individual requests for applications (RFAs) and program announcements (PA) may specify other requirements or expectations for data sharing that apply to specific projects.

<table>
<thead>
<tr>
<th>IC</th>
<th>Data Sharing Policy Name</th>
<th>Description of Data Sharing Policy</th>
<th>Repositories</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIH</td>
<td>NIH Data Sharing Policy</td>
<td>Expects investigators seeking more than $500K in direct support in any given year to submit a data sharing plan with their application or to indicate why data sharing is not possible.</td>
<td>No specific repository listed</td>
</tr>
<tr>
<td>NIH</td>
<td>NIH Policy on Deposit of Atomic Coordinates into Structural Databases</td>
<td>NIH policy requires that atomic coordinates from X-ray crystallographic and nuclear magnetic resonance experiments that were supported by NIH grants be deposited into the appropriate structural database at the time of submission of a research article deriving conclusions from these data.</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>NIGMS</td>
<td>ENCODE Consortia Data Release, Data Use, and Publication Policies</td>
<td>Requires resource producers to release primary data along with an initial interpretation, in the form of genome features, to the appropriate public databases as soon as the data is verified. Consortia members will also identify validation standards that will be applied in subsequent analyses of the data or with additional experimentation where appropriate. All data will be deposited to public databases, such as dbSNP or the ENCODE/modENCODE Data Coordination Centers (DCCs) and adhere pre-publication data will be made available for all to use.</td>
<td>ENCODE</td>
</tr>
<tr>
<td>NIH</td>
<td>Genomic Data Sharing Policy</td>
<td>Expects that large-scale genomic research data from NIH-funded studies involving human and non-human and model organisms, will be shared through a publicly available data repository. All studies with human genomic data should be registered in dbGaP, and the data should be submitted to an NIH-designated data repository. Non-human data may be submitted to any widely used data repository.</td>
<td>dbGaP (for registration), NIH-designated data repository (for data)</td>
</tr>
<tr>
<td>NIH</td>
<td>NIH Policy for Sharing of Data Obtained in NIH-Supported or Conducted Genomic Wide Association Studies (GWAS)</td>
<td>Expects all investigators who receive NIH support to conduct genome-wide analysis of genetic variation in a study population to submit to the NIH GWAS data repository descriptive information about their studies for inclusion in an open access portion of the NIH GWAS data repository. Strongly encourages the submission of curated and coded phenotypes, exposure, genotypes, and pedigree data, as appropriate, to the NIH GWAS data repository as soon as quality control procedures have been completed at the local institution. These detailed data will be made available through a controlled access process according to the dbGaP data access procedures.</td>
<td></td>
</tr>
<tr>
<td>NHLBI</td>
<td>NIH Bi Policy for Data Sharing from Clinical Trials and Epidemiological Studies</td>
<td>Encourages all applicants to include a plan to address data sharing or to state why data sharing is not possible. For studies that meet the following criteria, applicants are required to provide a data sharing plan, which will be reviewed and approved by the relevant NHLBI program official: a) research applications/proposals requesting $900,000 or more direct costs; b) research studies that have 500 or more participants or c) ancillary studies based on NHLBI-funded parent studies. Applications/proposals submitted in response to FOAs that specify inclusion of data sharing clauses, or other research studies deemed appropriate for data sharing by NHLBI program officials, should be submitted to dbGaP.</td>
<td>NHLBI data repository: BioLITCC</td>
</tr>
<tr>
<td>NIA</td>
<td>Alzheimer's Disease Genetics Sharing Plan</td>
<td>NIA policy in the area of human Alzheimer’s disease genetics applies to all NIA funded research in this area regardless of cost. NIA follows the NIH GWAS Policy and extends NIA’s existing policy on sharing data on Alzheimer’s disease genetics to include secondary analysis of data resulting from a genome wide association study. It is the policy of the NIA that useful data and Associated Phenotypic Data for the genetics of late onset Alzheimer’s disease be deposited at the National Cell Repository for Alzheimer’s Disease (NCRAD) whenever possible. It is the policy of the NIA that all Genetic Data derived from NIA funded studies for the genetics of late onset Alzheimer’s disease be deposited at the National Institute on Aging Genetics of Alzheimer’s Disease Data Storage Site (NIAGADS) or another NIA approved site or both whenever possible. It is the policy of the NIA that all GWAS data, including secondary analysis data, derived from NIA funded studies for the genetics of late onset Alzheimer’s disease be deposited at the NIH GWAS data repository (dbGaP) or another NIA approved site or both, whenever possible.</td>
<td>NCRAD: NIAGADS, dbGaP</td>
</tr>
</tbody>
</table>
Future of NLM/NIH Data & Access

- Federally funded research data sharing
  - Legislative activity
  - Mandated deposit
  - NLM/NIH managed repositories
- Issues
Data From the NLM

- Daniel Burgard
  - daniel.burgard@unthsc.edu
  - 817-735-2589

- Bench to bedside

- URL’s from presentation
  - https://nnlm.gov/scr
  - https://datascience.nih.gov/index