**Data Resources, Viewers   
(navigation, search, content interpretation, data accession)**

|  |
| --- |
| **Publicly Available Genomic Data Resources, Viewers**  Explore Content, Search |
| |  | | --- | | **NCBI Genome Data Viewer (GDV)**  <https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.38>  The NCBI Genome Data Viewer (GDV) is a genome browser supporting the exploration and analysis of eukaryotic RefSeq genome assemblies. Genome Data Viewer is also used by different NCBI resources, such as GEO and dbGaP, to display datasets associated with specified experiments or samples in a genome browser context. | | **Tutorials**  **NCBI Genome Data Viewer (Tutorial Page Functionality)**  <https://www.ncbi.nlm.nih.gov/genome/gdv/browser/help/#LAYOUT>  **NCBI Genome Data Viewer (Tutorial 11 videos Last updated Sep 23, 2020)**  <https://www.youtube.com/playlist?list=PLH-TjWpFfWruHgL0WRzZfQwp-MWzhIj16> |  |  | | --- | | **NCBI Gene Expression Omnibus (GEO)** <https://www.ncbi.nlm.nih.gov/geo/>  GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. | | **GEO DataSets (GDS)**  <https://www.ncbi.nlm.nih.gov/gds/> This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.  **About GEO DataSets** <https://www.ncbi.nlm.nih.gov/geo/info/datasets.html>  **GEO Profiles**  <https://www.ncbi.nlm.nih.gov/geoprofiles/> This database stores individual gene expression profiles from curated DataSets in the Gene Expression Omnibus (GEO) repository. Search for specific profiles of interest based on gene annotation or pre-computed profile characteristics.  **Querying GEO DataSets and GEO Profiles**  <https://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html> | | **Related Publications**  **The Gene Expression Omnibus database Methods Mol Biol. 2016; 1418: 93–110.**  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4944384/>  **NCBI GEO: mining millions of expression profiles database and tools Nucleic Acids Res. 2005 Jan 1**  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC539976/> |      |  | | --- | | **NCBI Sequence Read Archive (SRA)**  <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>?  SRA is NIH primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC). | | Stores raw sequence data from "next-generation" sequencing technologies including  Illumina, 454, IonTorrent, Complete Genomics, PacBio and OxfordNanopores.  In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence.  **SRA Toolkit**  <https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit>  The SRA Toolkit provides 64-bit binary installations for Linux distributions, Mac OS X, Windows.  **SRA Explorer**  tool aims to make datasets within the Sequence Read Archive more accessible.  <https://sra-explorer.info/> |  |  | | --- | | **ENCODE: Encyclopedia of DNA Elements** <https://www.encodeproject.org/> |  |  | | --- | | **1000 Genomes Project**  <http://www.1000genomes.org> | | **1000 Genomes Project FAQ**  If you have any other questions you can’t find the answer to please email [info@1000genomes.org](mailto:info@1000genomes.org)  <https://www.internationalgenome.org/faq>  **IGSR: The International Genome Sample Resource**  Open human variation data  <https://www.internationalgenome.org/home>  **1000 Genomes 30x on GRCh38**  <https://www.internationalgenome.org/data-portal/data-collection/30x-grch38>  **1000 Genomes Browser (Phase 3)**  <https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/>  **Ensembl**  <http://uswest.ensembl.org/Homo_sapiens/Info/Index>   |  | | --- | | **Data Portal (Search by Population, Technology, Data Collection)**  <https://www.internationalgenome.org/data-portal/sample>  Technology  Exome  Low coverage WGS  Integrated variant call sets  File Format  FASTQ  VCF | |  |  | | --- | | **dbGaP**  **db Genotypes and Phenotypes**  <https://dbgap.ncbi.nlm.nih.gov/> |  |  | | --- | | **NCBI dbSNP database**  <https://www.ncbi.nlm.nih.gov/snp/> | | dbSNP contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.  **dbSNP Data Access:**  <https://www.ncbi.nlm.nih.gov/snp/docs/RefSNP_about/#data-access>  RefSNP data, including genotype, frequency and associated metadata, are available without restrictions on the web, FTP, and API.  **dbSNP Overview:**  <https://www.ncbi.nlm.nih.gov/books/NBK21088/>  **dbSNP Tutorials on GitHub:**  <https://www.ncbi.nlm.nih.gov/snp/> |  |  | | --- | | **NIH Data Sharing Repositories/**  Open Domain-Specific Data Sharing Repositories  <https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html> |  |  | | --- | | **GenBank**  <https://www.ncbi.nlm.nih.gov/genbank/>  GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. | | GenBank is part of the **International Nucleotide Sequence Database Collaboration**, which comprises the DNA DataBank of  Japan (DDBJ), European Nucleotide Archive (ENA), GenBank at NCBI.  Ways to search and retrieve data from GenBank:  **-** Search GenBank for sequence identifiers and annotations with Entrez Nucleotide.  **-** Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool).  BLAST searches CoreNucleotide, dbEST, and dbGSS independently.  **-** Search, link, and download sequences programatically using NCBI e-utilities. |  |  | | --- | | **NCI / Genetic Data Commons (GDC)**  <https://gdc.cancer.gov/>  <https://portal.gdc.cancer.gov/> | | **GDC Data Transfer Tool: An Overview**  <https://docs.gdc.cancer.gov/Data_Transfer_Tool/Users_Guide/Getting_Started/>  Raw sequence data, stored as BAM files, make up the bulk of data stored at the NCI Genomic Data Commons (GDC).  **GDC Exploration**  <https://docs.gdc.cancer.gov/Data_Portal/Users_Guide/Exploration/>  The Exploration Page allows users to explore data in the GDC using advanced filters/facets, which includes those on a gene and mutation level. Users choose filters on specific Cases, Genes, and/or Mutations on the left of this page and then can visualize these results on the right.  **GDC Data Transfer Tool**  <https://docs.gdc.cancer.gov/Data_Transfer_Tool/Users_Guide/Preparing_for_Data_Download_and_Upload/>  is intended to be used in conjunction with the GDC Data Portal and the GDC Data Submission Portal to transfer data to or from the GDC. The GDC Data Portal's interface is used to generate a manifest file or obtain UUID(s) and (for Controlled-Access Data) an authentication token. The GDC Data Transfer Tool is then used to transfer the data files listed in the manifest file or identified by UUID(s). |  |  | | --- | | **GDC/TCGA - Genomic Data Commons (GDC) / The Cancer Genome Atlas (TCGA)**  <https://tcga-data.nci.nih.gov/tcga/>  Harmonized Cancer Datasets | | In order to download data from TCGA data portal:  1. Connect to <https://tcga-data.nci.nih.gov/tcga/>  2. Select the cancer subtype you are interested in (i.e breast invasive carcinoma)  3. Select mRNA  4. Now you can see a table where rows are representing different patients.  5. If present select the column (by clicking on header) that referse to RNASeq or RNASeqV2 if it is present for that  cancer subtype and then click BUILD archive.  6. Keep in mind that just below the header there is a number indicating the respective data level. Levels 1-4  <https://wiki.nci.nih.gov/display/TCGA/Data+level>  If you need RAW data such as FASTQ files you have find level 1 data, but often this kind of data is not publicly  available on TCGA and you might need to ask for permission in order to download it.  Marco has listed the steps to access open tier data on TCGA. In case you are interested in accessing lower level data,  such as raw bam files for rna seq samples, you can apply for the access here  <https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login>  **GDC for TCGA Data Access Matrix Users**  <https://gdc.cancer.gov/gdc-tcga-data-access-matrix-users>  **Resources for TCGA Users**  <https://gdc.cancer.gov/resources-tcga-users>  **The Cancer Genome Atlas (TCGA)**  <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga> |  |  | | --- | | **UCSC Genome Bioinformatics Group**  <https://genome-euro.ucsc.edu/index.html> |  |  | | --- | | **IGSR**: The International Genome Sample Resource  Supporting open human variation data  <https://www.internationalgenome.org/> |  |  | | --- | | **AnVIL**  supports the management, analysis and sharing of human disease data for the research community  <https://anvilproject.org/>  <https://gen3.theanvil.io/>  <https://gen3.theanvil.io/explorer> |  |  | | --- | | **REST APIs** <http://MyGene.info>  <http://MyVariant.info>  <http://BioThings.io>  <http://data.cvisb.org/home>  <http://Smart-API.info> | |

|  |
| --- |
| **Google**  **how to analyze geo datasets in r**  <https://www.google.com/search?lr=&as_qdr=all&sxsrf=ALeKk03S9orXF4hAyQbXO282dDPFBUHzAA:1609876679459&q=how+to+analyze+geo+datasets+in+r&sa=X&ved=2ahUKEwjyuvH8yYXuAhUOQ80KHZVNBEgQ1QIoAHoECAgQAQ> |
| **Exploratory analysis of datasets obtained from GEO**  <http://www.biotechworld.it/bioinf/2015/10/22/exploratory-analysis-of-datasets-obtained-from-geo/>  bioinformatics > Exploratory analysis of datasets obtained from GEO  October 22, 2015  **Using the GEOquery Package**  <https://bioconductor.org/packages/release/bioc/vignettes/GEOquery/inst/doc/GEOquery.html>  September 21, 2014  **Retrieve and analyze a gene expression data set from NCBI GEO in R**  <https://www.youtube.com/watch?v=gC-WuK4BbJY>  Nov 2, 2012  **Analysing data from GEO - Work in Progress**  <https://sbc.shef.ac.uk/geo_tutorial/tutorial.nb.html>  30 Jun 2020 |

|  |
| --- |
| **Registry of Open Data on AWS**  <https://registry.opendata.aws/> currently 205 datasets |
| **See**  [**Usage Examples/Tutorials for datasets listed in this registry**](https://registry.opendata.aws/usage-examples)   |  | | --- | | **Encyclopedia of DNA Elements (ENCODE)**  <https://registry.opendata.aws/encode-project/> | | Description  The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. ENCODE investigators employ a variety of assays and methods to identify functional elements. The discovery and annotation of gene elements is accomplished primarily by sequencing a diverse range of RNA sources, comparative genomics, integrative bioinformatic methods, and human curation. Regulatory elements are typically investigated through DNA hypersensitivity assays, assays of DNA methylation, and immunoprecipitation (IP) of proteins that interact with DNA and RNA, i.e., modified histones, transcription factors, chromatin regulators, and RNA-binding proteins, followed by sequencing.  **Usage Examples/Tutorials**  **Exploring ENCODE data from EC2 with Jupyter notebook** Excellent Detailed instructions  <https://github.com/ENCODE-DCC/encode-data-usage-examples/blob/master/mount_s3_bucket_and_run_jupyter_on_ec2.ipynb>  This notebook demonstrates how to mount s3://encode-public on an EC2 instance using Goofys.  Goofys makes an S3 bucket appear as a typical UNIX file system, and is useful for tools that expect a local file path.  Once the bucket is mounted we can launch a Jupyter notebook on the instance and connect to it remotely.  The benefit of using EC2 is that the compute is scalable to the analysis you would like to perform, and you don't have to download anything locally.  **Ingesting ENCODE data into TileDB with S3 backend**  <https://github.com/ENCODE-DCC/encode-data-usage-examples/blob/master/ingest_encode_data_tile_db_with_s3_backend.ipynb>  **Publications**  **New developments on the Encyclopedia of DNA Elements (ENCODE) data portal**  <https://academic.oup.com/nar/article/48/D1/D882/5622708#190992715>  Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D882–D889  12 November 2019 |  |  | | --- | | **Genome Aggregation Database (gnomAD)**  <https://registry.opendata.aws/broad-gnomad/> | | **Description**  The Genome Aggregation Database (gnomAD) is a resource developed by an international coalition of investigators that aggregates and harmonizes both exome and genome data from a wide range of large-scale human sequencing projects. The summary data provided here are released for the benefit of the wider scientific community without restriction on use. The v2 data set (GRCh37) spans 125,748 exome sequences and 15,708 whole-genome sequences from unrelated individuals. The v3 data set (GRCh38) spans 71,702 genomes, selected as in v2. Sign up for the gnomAD mailing list here.  Resources on AWS  Description  gnomAD summary data aggregated from large-scale human genome and exome sequencing projects.  Resource type  S3 Bucket  **Amazon Resource Name (ARN)**  **arn:aws:s3:::gnomad-public-us-east-1**  AWS Region  us-east-1  AWS CLI Access (No AWS account required)  aws s3 ls s3://gnomad-public-us-east-1/ --no-sign-request | | **gnomAD browser**  Downloads  <https://gnomad.broadinstitute.org/downloads/>  gnomAD data is available for download through Google Cloud Public Datasets, the Registry of Open Data on AWS, and Azure Open Datasets. We recommend using Hail and our Hail utilities for gnomAD to work with the data.  In addition to the files listed below, gnomAD variants are available as a BigQuery dataset and Terra has a demo workspace for working with gnomAD data.  **Registry of Open Data on AWS**  Files can be browsed and downloaded using the AWS Command Line Interface.  **aws s3 ls s3://gnomad-public-us-east-1/release/** | | **Example use of the big data**  **A structural variation reference for medical and population genetics**  Nature volume 581, pages444–451(2020)  31 Citations  <https://www.nature.com/articles/s41586-020-2287-8> **Data availability** All gnomAD-SV site-frequency data for appropriately consented samples (*n* = 10,847) have been distributed in VCF and BED format via the gnomAD browser (<https://gnomad.broadinstitute.org/downloads/>), as well as from NCBI dbVar under accession nstd166. Furthermore, these SVs have been integrated directly into the gnomAD browser[8](https://www.nature.com/articles/s41586-020-2287-8#ref-CR8). The architecture of the gnomAD browser is described in the main gnomAD study[4](https://www.nature.com/articles/s41586-020-2287-8#ref-CR4), as well as instructions for how to access and query the data hosted therein.  **Code availability**  The gnomAD-SV discovery pipeline is publicly available via a series of methods configured for the FireCloud/Terra platform (<https://portal.firecloud.org/#methods>) under the methods namespace ‘Talkowski-SV’. The svtk software package used extensively in the gnomAD-SV discovery pipeline is publicly available via GitHub (<https://github.com/talkowski-lab/svtk>). Most custom scripts used in the production and/or analysis of the gnomAD-SV dataset are publicly available via GitHub (<https://github.com/talkowski-lab/gnomad-sv-pipeline>). All code is made available under the MIT license, unless stated otherwise. | | **gnomAD browser**  **Downloads**  <https://gnomad.broadinstitute.org/downloads/>  gnomAD data is available for download through Google Cloud Public Datasets, the Registry of Open Data on AWS, and Azure Open Datasets.  We recommend using Hail and our Hail utilities for gnomAD to work with the data.  In addition to the files listed below, gnomAD variants are available as a BigQuery dataset and Terra has a demo workspace for working with gnomAD data.  Registry of Open Data on AWS  Files can be browsed and downloaded using the AWS Command Line Interface.  **aws s3 ls s3://gnomad-public-us-east-1/release/**  **All chromosomes sites VCF**  Download Google / Amazon / Microsoft  **Exomes**  chr1 sites VCF 5.77 GiB  chr2 sites VCF 4.2 GiB  chr3 sites VCF 3.29 GiB  chr4 sites VCF 2.17 GiB  chr5 sites VCF 2.51 GiB  chr6 sites VCF 2.83 GiB  chr7 sites VCF 2.88 GiB  chr8 sites VCF 2.13 GiB  chr9 sites VCF 2.4 GiB  chr10 sites VCF 2.23 GiB  chr11 sites VCF 3.61 GiB  chr12 sites VCF 3.07 GiB  chr13 sites VCF 976.73 MiB  chr14 sites VCF 2.02 GiB  chr15 sites VCF 2.08 GiB  chr16 sites VCF 3.04 GiB  chr17 sites VCF 3.62 GiB  chr18 sites VCF 882.14 MiB  chr19 sites VCF 4.3 GiB  chr20 sites VCF 1.44 GiB  chr21 sites VCF 652.73 MiB  chr22 sites VCF 1.43 GiB  chrX sites VCF 1.33 GiB  chrY sites VCF 15.66 MiB  **Genomes**  -  -  - | | **Constraint**  <https://gnomad.broadinstitute.org/downloads/#v2-constraint>  The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020).  **Multi-nucleotide variants (MNVs)**  <https://gnomad.broadinstitute.org/downloads/#v2-multi-nucleotide-variants>  Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nature Communications 11, 2539 (2020).  **Proportion expressed across transcripts (pext)**  <https://gnomad.broadinstitute.org/downloads/#v2-pext>  Transcript expression-aware annotation improves rare variant interpretation. Nature 581, 452–458 (2020) | | **Related Papers**  **The mutational constraint spectrum quantified from variation in 141,456 humans**  https://www.nature.com/articles/s41586-020-2308-7  Nature volume 581, pages434–443(2020)  300 Citations  **Transcript expression-aware annotation improves rare variant interpretation**  https://www.nature.com/articles/s41586-020-2329-2  27 May 2020  **Thousands of human sequences provide deep insight into single genomes**  A massive genome-sequencing and analysis effort has produced the most comprehensive sets of data and tools for understanding human genetic variation so far. The resource will be invaluable to biologists of every stripe.  **A structural variation reference for medical and population genetics**  Nature volume 581, pages444–451(2020)  31 Citations |  |  | | --- | | **COVID-19 Data Lake**  https://registry.opendata.aws/aws-covid19-lake/ | |  | |

**AWS Labs**

<https://github.com/awslabs>

**AWS Labs Machine Learning**

<https://github.com/awslabs/dgl-ke>

|  |
| --- |
| **Medical Data for Machine Learning**  **Credit to Andrew L. Beam, et al**  Assistant Professor, Harvard School of Public Health  <https://github.com/beamandrew/>  <https://github.com/beamandrew/medical-data>  This is a curated list of medical data for machine learning.  This list is provided for informational purposes only, please make sure you respect any and all usage restrictions for any of the data listed. |
| **3. Data derived from Electronic Health Records (EHRs)**  Building the graph of medicine from millions of clinical narratives  Co-occurence statistics for medical terms extracted from 14 million clinical notes and 260,000 patients.  Paper: <http://www.nature.com/articles/sdata201432>  Data: <http://datadryad.org/resource/doi:10.5061/dryad.jp917>  **Learning Low-Dimensional Representations of Medical Concept**  Low-dimensional embeddings of medical concepts constructed using claims data. Note that this paper utilizes data from Building the graph of medicine from millions of clinical narratives  Paper: <http://cs.nyu.edu/~dsontag/papers/ChoiChiuSontag_AMIA_CRI16.pdf>  Data: <https://github.com/clinicalml/embeddings>  **MIMIC-III, a freely accessible critical care database**  Anonymized critical care EHR database on 38,597 patients and 53,423 ICU admissions. Requires registration.  Paper: <http://www.nature.com/articles/sdata201635>  Data: <http://physionet.org/physiobank/database/mimic3cdb/>  Clinical Concept Embeddings Learned from Massive Sources of Medical Data  Embeddings for 108,477 medical concepts learned from 60 million patients, 1.7 million journal articles, and clinical notes of 20 million patients  Paper: <https://arxiv.org/abs/1804.01486>  Embeddings: <https://figshare.com/s/00d69861786cd0156d81>  Interactive tool: <http://cui2vec.dbmi.hms.harvard.edu>  Evaluation of Embeddings of Laboratory Test Codes for Patients at a Cancer Center  200 dimensional Word2Vec embeddings of 1098 laboratory test codes (LOINCs) trained from 8,280,238 lab orders for 79,081 patients at City of Hope National Medical Center (Los Angeles, CA).  Paper: <https://arxiv.org/abs/1907.09600>  Embeddings and Code: <https://github.com/elleros/DSHealth2019_loinc_embeddings>  **4. National Healthcare Data**  Centers for Disease Control and Prevention (CDC)  Data from the CDC on many areas, including:  Biomonitoring  Child Vaccinations  Flu Vaccinations  Health Statistics  Injury & Violence  MMWR  Motor Vehicle  NCHS  NNDSS  Pregnancy & Vaccination  STDs  Smoking & Tobacco Use  Teen Vaccinations  Traumatic Brain Injury  Vaccinations  Web Metrics  Landing page: <https://data.cdc.gov>  Data Catalog: <https://data.cdc.gov/browse>  **Medicare Data**  Data from the Centers for Medicare & Medicaid Services (CMS) on hospitals, nursing homes, physicians, home healthcare, dialysis, and device providers.  Landing page: <https://data.medicare.gov>  Explorer: <https://data.medicare.gov/data>  **Texas Public Use Inpatient Data File Data on 11 Million inpatient visits with diagnosis, procedure codes and outcomes from Texas**  **between 2006 & 2009.**  Link: <https://www.dshs.texas.gov/thcic/hospitals/Inpatientpudf.shtm>  **Dollars for Doctors**  Propublica investigation of money paid by pharmaceutical companies to doctors.  Information: <https://www.propublica.org/series/dollars-for-docs>  Search tool: <https://projects.propublica.org/docdollars/>  Data request: <https://projects.propublica.org/data-store/sets/health-d4d-national-2>  **DocGraph Physician interaction network obtained through a freedom of information act request. Covers nearly 1 million entities.**  Main page: <http://www.docgraph.com>  Information: <http://thehealthcareblog.com/blog/2012/11/05/tracking-the-social-doctor-opening-up-physician-referral-data-and-much-more/>  Data: <http://linea.docgraph.org>  **5. UCI Datasets**  **Liver Disorders Data Set**  Data on 345 patients with and without liver disease. Features are 5 blood biomarkers thought to be involved with liver disease.  Data: <https://archive.ics.uci.edu/ml/datasets/Liver+Disorders>  **Thyroid Disease Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease>  **Breast Cancer Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>  **Heart Disease Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>  **Lymphography Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Lymphography>  **Parkinsons Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/parkinsons>  **Parkinsons Telemonitoring Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring>  **Parkinson Speech Dataset with Multiple Types of Sound Recordings Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings>  **Parkinson's Disease Classification Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification>  Primary Tumor Dataset Data: <https://archive.ics.uci.edu/ml/datasets/primary+tumor>  **6. Biomedical Literature**  **PMC Open Access Subset**  Collection of all the full-text, open access articles in Pubmed central.  Information: <http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>  Archived files: <http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/#Data_Mining>  **PubMed 200k RCT**  Collection of pubmed abstracts from randomized control trials (RCTs). Annotations for each sentence in the abstract are available.  Paper: <https://arxiv.org/abs/1710.06071>  Data: <https://github.com/Franck-Dernoncourt/pubmed-rct>  **Web API of PubMed Articles**  NLM also provided Web API for accessing biomedical literatures in PubMed.  Instructions for getting PubMed articles: <https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PubMed/>  (not full text, just title, abstract, etc.)  For articles in PubMed Central, instructions for getting the whole articles: https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PMC/  **EBM NLP**  Collection of pubmed abstracts from randomized control trials (RCTs). Annotation of Population, Intervention, and Outcomes (PICO elements) are available.  Paper: <https://arxiv.org/abs/1806.04185>  Data: <https://ebm-nlp.herokuapp.com/annotations>  Website: <https://ebm-nlp.herokuapp.com/index>  **Evidence Inference**  A dataset for inferring the results of randomized control trials (RCTs). A collection of pubmed RCTs from the open access subset. Annotations of (intervention, comparison intervention, outcome, significance finding, evidence span) are available.  Paper: <https://arxiv.org/abs/1904.01606>  Data: <https://github.com/jayded/evidence-inference/tree/master/annotations>  Website: <http://evidence-inference.ebm-nlp.com/>  **PubMedQA**  A dataset for biomedical research question answering. Task is to use yes/no/maybe to answer naturally occuring questions in PubMed titles.  Paper: <https://arxiv.org/abs/1909.06146>  Data: <https://github.com/pubmedqa/pubmedqa>  Website: <https://pubmedqa.github.io/>  **7. TREC Precision Medicine / Clinical Decision Support Track**  Text REtrieval Conference (TREC) is running a track on Precision Medicine / Clinical Decision Support from 2014.  **2014 Clinical Decision Support Track**  Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records.  Information and Data: <http://www.trec-cds.org/2014.html>  **2015 Clinical Decision Support Track**  Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records.  Information and Data: <http://www.trec-cds.org/2015.html>  **2016 Clinical Decision Support Track**  Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records. Actual electronic health record (EHR) patient records are be used instead of synthetic cases.  Information and Data: <http://www.trec-cds.org/2016.html>  **2017 Clinical Decision Support Track**  Focus: Retrieve useful precision medicine-related information to clinicians treating cancer patients.  Information and Data: <http://www.trec-cds.org/2017.html> |

**NCI / Genetic Data Commons (GDC)**

<https://gdc.cancer.gov/>

<https://portal.gdc.cancer.gov/>

**Genome Data Viewer**

<https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.39>

**Variation Viewer**

<https://www.ncbi.nlm.nih.gov/variation/view/>

**NCBI Sequence Viewer**

<https://www.ncbi.nlm.nih.gov/projects/sviewer/>

|  |
| --- |
| **NCBI Reference Sequence Database (RefSeq)**  <https://www.ncbi.nlm.nih.gov/refseq/>  <https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/Homo_sapiens/> |

**Publicly available data:**

GEO NCBI, ENA, SRA

other useful data that can be used in bioinformatic analysis namely; reference genomes, annotations.

Testing differential gene expression

Whole-genome bisulfite sequencing data analysis

Whole-exome sequencing data analysis

Whole-genome sequencing data analysis

Microbiome data analysis