ANSO
Alliance of International Science Organizations

Highlight For Open Data

ANSO Highlight for Open Data is a new series to encourage and support the sharing of scientific data under the ANSO platform by introducing regional and global Scientific Data Centers and their high-quality datasets for free and open download. This publication also aims to deliver open data information to end-users outside the research community, to serve the Open Science Initiative for all stakeholders interested in regional and global sustainable development.

Special Issue
October 2022
Issue.04

Recommendation for ANSO Highlight for Open Data from ANSO President Prof. BAI Chunli

October 17, 2022

Dear ANSO Members,

I take much pleasure to write to you and hope that all of you are doing fine, and so are your organizations. The world is undergoing dramatic changes as it has entered a new age of rapid development and large-scale transformations. These fundamental processes are accompanied by ever-growing connectivity and accelerating informatization and digitalization. Therefore, new approaches are needed to promote wide and effective international cooperation to address today's problems and challenges for sustainable development.

In this regard and in response to the UNESCO Recommendation on Open Science, I recommend ANSO Highlight for Open Data, a new publication of ANSO, with the aim to encourage and support the sharing of scientific data under the ANSO platform. In this issue, we introduce the main databases from one of the ANSO cooperative networks, the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB).

Biodiversity and health big data are of critical importance for sustainable development and human healthcare. As a major global biological data center, CNCB-NGDC advances life and health sciences by developing dozens of biological data resources and tools, including those for COVID-19. With over 17 PB of omics data submitted by more than 3,000 researchers from 18 countries/regions, NGDC provides open access to millions of users worldwide, and with more than 8 billion data downloads. All these resources and their related services are publicly accessible at https://ngdc.cncb.ac.cn

We believe these resources will greatly promote knowledge exchanges and data sharing among the ANSO member organizations and the global research communities for the benefit of scientific advances and the achievement of UN SDGs.

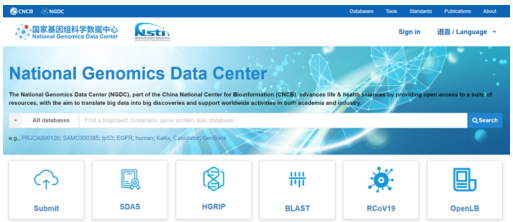
It is much appreciated if you can share the information on your web page for wide publicity. If you and your colleagues are interested in the information and wish to enhance international collaboration in open science data, please feel free to contact ANSO Secretariat for further assistance or write directly to Prof. BAO Yiming at baoym@big.ac.cn, Director of CNCB-NGDC. We also highly welcome recommendations from you of representative data centers and resources from your side and your country to create linkages and partnerships with CNCB-NGDC. Enclosed please find this issue of the ANSO Highlight for Open Data.

Thank you for your attention to the above information and with my kindest regards,

BAI Chunli, Ph.D.
ANSO President

Add: No. 16 Lincui Road, Chaoyang District, Beijing, China
Postcode: 100101
Email: anso-public@anso.org.cn Tel: +86-10-84097121

National Genomics Data Center, China National Center for Bioinformation

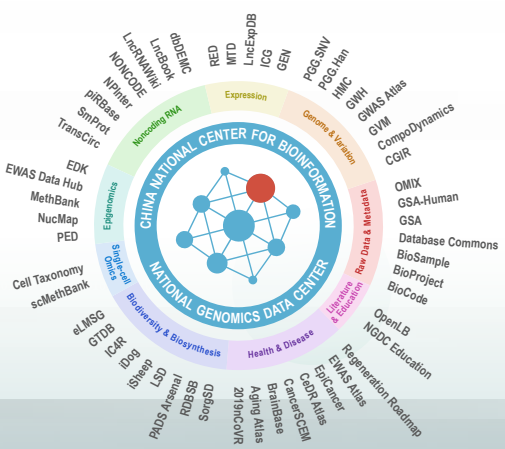


The National Genomics Data Center (NGDC), part of the China National Center for Bioinformation (CNCB), advances life and health sciences by providing open access to a suite of resources, with the aim of translating big data into big discoveries and supporting worldwide activities in both academia and industry. NGDC was officially announced by the Ministry of Science & Technology and the Ministry of Finance of China on June 5, 2019. It is based in the Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS) / China National Center for Bioinformation (CNCB), in collaboration with the Institute of Biophysics (IBP), CAS and the Shanghai Institute of Nutrition and Health (SINH), CAS, as well as a number of other partners.

With the explosive growth of multi-omics data, CNCB-NGDC is constantly scaling up and updating its core database resources through big data archive, curation, integration and analysis, including the Genome Sequence Archive (GSA), Genome Warehouse (GWH), Genome Variation Map (GVM), Resource for Coronavirus 2019 (RCoV19). Moreover, BIG Search, a scalable text search engine, provides easy access to over 200 internal and external biological resources from CNCB-NGDC and partners, as well as the European Bioinformatics Institute (EBI) and the U.S. National Center for Biotechnology Information (NCBI). All these resources along with their services are publicly accessible at https://ngdc.cncb.ac.cn.

CNCB-NGDC provides free, fast and convenient multi-omics data submission and archive services for users from China and abroad. As of October 10, 2022, GSA has reached 17 PB of omics data, submitted by 3,154 researchers from 712 institutions. Data submissions have been reported in 1,867 articles covering 447 journals. GSA has been designated as a supported data repository by Springer Nature and Elsevier, and also recognized by Wiley, Taylor & Francis, and Cell Press. In addition, the Global Biodiversity and Health Big Data Alliance (BHBD) was launched in 2018 and now has 28 members from 12 countries, with support from ANSO and IUBS. CNCB-NGDC has been acknowledged by international researchers with increasing influence, and is listed as one of the major database providers together with EBI and NCBI by the journal Nucleic Acids Research.

https://ngdc.cncb.ac.cn
Tel: +86-10-84097298



Core database resources of CNCB-NGDC
(Nucleic Acids Res 2022)

Main Databases of NGDC

1. Resource for Coronavirus 2019 (RCoV19)

Introduction of the data

RCoV19 is an open-access information resource on the novel coronavirus, which has been updated daily since its release in January 2020. RCoV19 features comprehensive integration of genomic and proteomic sequences as well as their metadata information from GISAID, NCBI, NMDC and CNCB-NGDC. It also incorporates a wide range of relevant information including scientific publications, news, and popular articles for science dissemination, and provides visualization functionalities for genome variation analysis results based on all collected SARS-CoV-2 strains. RCoV19 has archived 13 million SARS-CoV-2 sequences, accessed by 1.8 millions of users from 181 countries/regions with downloads of over 8 billion sequences.

Format of the data, total size of the data

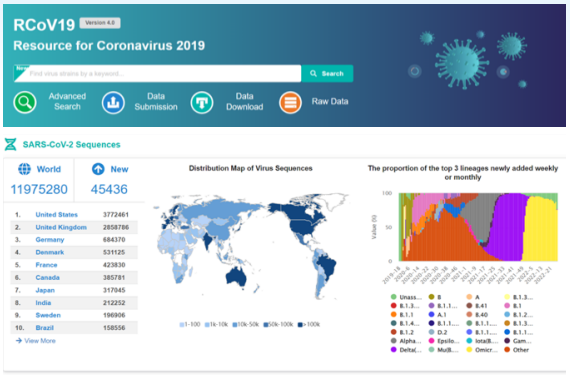
Metadata format: tsv, zip; size: 190 KB
Variant data format: vcf.gz; size: 514 MB
Sequence format: fasta; size: 268.1 GB

Download site and contact person

https://ngdc.cncb.ac.cn/ncov
Contact: gwh@big.ac.cn; gvm@big.ac.cn

Related papers and references

- The 2019 novel coronavirus resource. Yi Chuan 2020.
- An online coronavirus analysis platform from the National Genomics Data Center. Zool Res 2020.
- The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR. Genomics Proteomics Bioinformatics 2020.



Homepage of RCoV19

2. Genome Sequence Archive (GSA)

Introduction of the data

GSA is a public data repository for collecting, archiving, managing and sharing raw sequence data, which is the first repository of the genome sequence data with international journal recognition in China. GSA has been designated as a supported data repository by Springer Nature and Elsevier. It is also one of the registered repositories in FAIRsharing, and is supported by Wiley and Taylor & Francis. GSA has collected all the metadata of the genome sequence data in the International Nucleotide Sequence Database Collaboration (INSDC), and provides global search and hot data download services.

Format of the data, total size of the data

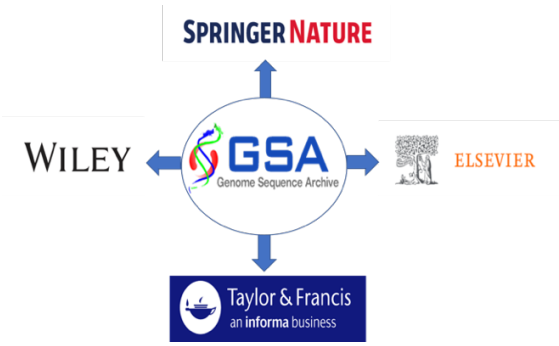
Format: fastq, bam
Data size: 17 PB

Download site and contact person

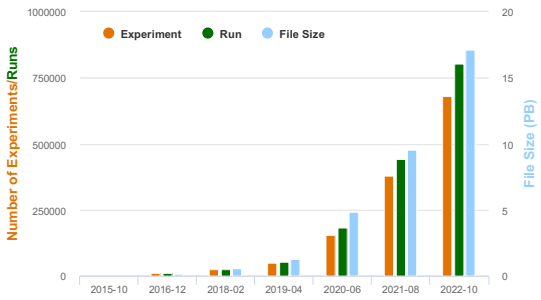
https://ngdc.cncb.ac.cn/gsa
Contact: gsa@big.ac.cn

Related papers and references

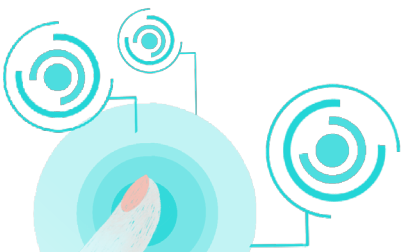
- The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021.
- GSA: Genome Sequence Archive. Genomics Proteomics Bioinformatics 2017.



GSA is recognized by international publishers



GSA Data Growth-20221010



3. Open Archive for Miscellaneous Data (OMIX)

Introduction of the data

OMIX is a new archive that aims to meet users' needs for submitting various types of data other than sequences. It collects not only raw data from transcriptome, epigenome, and microarray, but also functional data such as lipidome, metabolome, proteome, and other types of scientific data related to research.

Format of the data, total size of the data

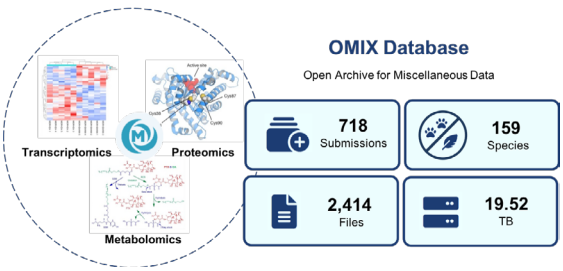
Format: txt, xls, raw, dcm, zip, tar, etc.
Data size: 19.52 TB

Download site and contact person

https://ngdc.cncb.ac.cn/omix
Contact: gsa@big.ac.cn

Related papers and references

- The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021.



Data resources of OMIX

4. Genome Warehouse (GWH)

Introduction of the data

GWH is a public repository housing genome-scale data for a wide range of species, delivering a series of web services for genome data submission, storage, release and sharing. It provides important data resources for comparative genome analysis, molecular breeding and other researches.

Format of the data, total size of the data

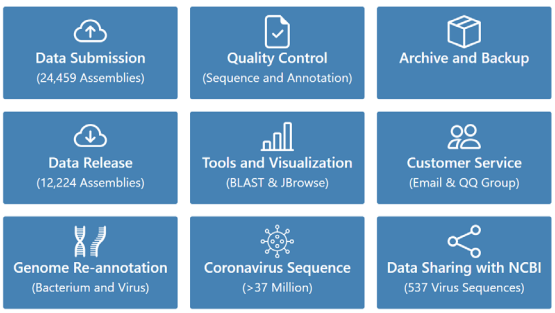
Format: fasta, gff
Data size: 1 TB

Download site and contact person

https://ngdc.cncb.ac.cn/gwh
Contact: gwh@big.ac.cn

Related papers and references

- Genome Warehouse: a public repository housing genome-scale data. Genomics Proteomics Bioinformatics 2021.



Feature functions of GWH

5. Genome Variation Map (GVM)

Introduction of the data

GVM is a public data repository of genome variations, including single nucleotide polymorphisms and small insertions and deletions, with particular focuses on human as well as cultivated plants and domesticated animals.

Format of the data, total size of the data

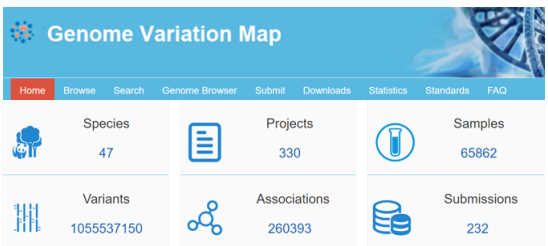
Format: vcf, gvcf
Data size: 39.64 TB

Download site and contact person

https://ngdc.cncb.ac.cn/gvm
Contact: gvm@big.ac.cn

Related papers and references

- Genome Variation Map: a worldwide collection of genome variations across multiple species. Nucleic Acids Res 2021.
- Genome Variation Map: a data repository of genome variations in BIG Data Center. Nucleic Acids Res 2018.



Data resources of GVM

6. Gene Expression Nebulas (GEN)

Introduction of the data

GEN is a comprehensive data portal integrating transcriptomic profiles across multiple species at both bulk and single-cell levels, which are critical for unravelling both transcriptional and post-transcriptional regulatory mechanisms. GEN adopts a structured curation model to categorize diverse experimental conditions into different biological contexts, provides abundant gene annotations based on value-added curation of transcriptomic profiles, and delivers online and offline services for bulk and single-cell data analysis and visualization.

Format of the data, total size of the data

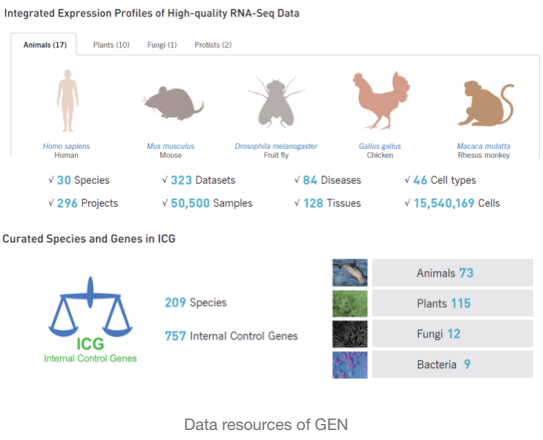
Format: tsv, mtx, rds, xml
Data size: 211 GB

Download site and contact person

https://ngdc.cncb.ac.cn/gen
Contact: gen@big.ac.cn

Related papers and references

- Gene Expression Nebulas (GEN): a comprehensive data portal integrating transcriptomic profiles across multiple species at both bulk and single-cell levels. Nucleic Acids Res 2021.
- ICG: a wiki-driven knowledgebase of internal control genes for RT-qPCR normalization. Nucleic Acids Res 2018.
- Rice Expression Database (RED): an integrated RNA-Seq-derived gene expression database for rice. J Genet Genomics 2017.



Data resources of GEN

7. Methylation Bank (MethBank)

Introduction of the data

MethBank is a database that integrates high-quality DNA methylomes across a variety of species. It is equipped with user-friendly web interfaces for data presentation, search and visualization. MethBank offers not only gene methylation profiles but also online analysis tools.

Format of the data, total size of the data

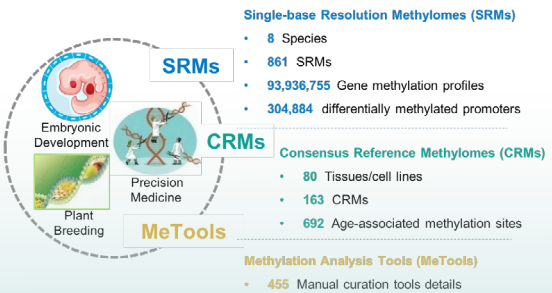
Format: bed, wig
Data size: 43.4 TB

Download site and contact person

https://ngdc.cncb.ac.cn/methbank
Contact: methbank@big.ac.cn

Related papers and references

- MethBank 3.0: a database of DNA methylomes across a variety of species. Nucleic Acids Res 2018.
- MethBank: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data. Nucleic Acids Res 2015.



Data resources of MethBank

8. GWAS Atlas

Introduction of the data

GWAS Atlas is a manually curated resource of genome-wide variant-trait associations for a wide range of species. It integrates high-quality curated genome-wide associations for animals and plants and provides user-friendly web interfaces for data browsing and downloading, accordingly serving as a valuable resource for genetic research of important traits and breeding application.

Format of the data, total size of the data

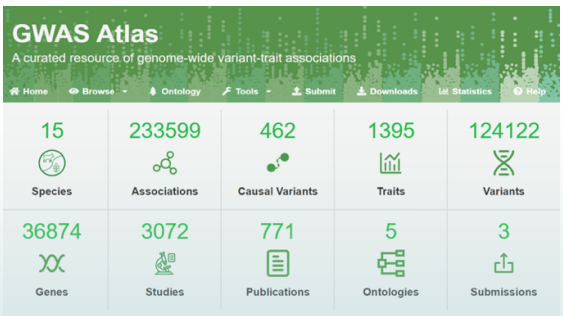
Format: txt, obo
Data size: 6.3 MB

Download site and contact person

https://ngdc.cncb.ac.cn/gwas
Contact: gwas@big.ac.cn

Related papers and references

GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals. Nucleic Acids Res 2020.



Data resources of GWAS Atlas

9. LncExpDB

Introduction of the data

LncExpDB is a comprehensive database for long non-coding RNA (lncRNA) expression. It covers expression profiles of lncRNA genes across various biological contexts, predicts potential functional lncRNAs and their interacting partners, and thus provides essential guidance on experimental design.

Format of the data, total size of the data

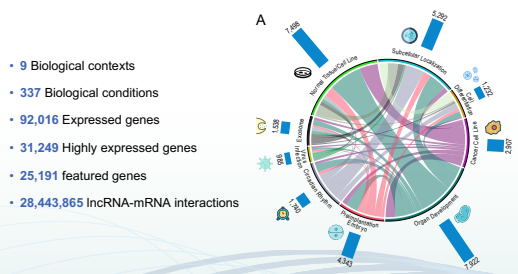
Format: csv, tar.gz
Data size: 24.61 GB

Download site and contact person

https://ngdc.cncb.ac.cn/lncexpdb
Contact: lncwiki@big.ac.cn

Related papers and references

LncExpDB: an expression database of human long non-coding RNAs. Nucleic Acids Res 2021.



Data resources of LncExpDB

10. LncBook

Introduction of the data

LncBook provides a comprehensive and high-quality list of human lncRNAs, enriches these lncRNAs with essential multi-omics signatures, and identifies featured lncRNAs in diseases and diverse biological contexts.

Format of the data, total size of the data

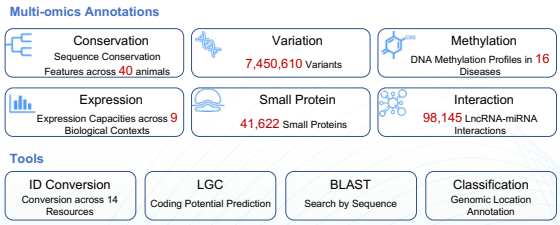
Format: gtf, csv, tar.gz, zip
Data size: 0.52 GB

Download site and contact person

https://ngdc.cncb.ac.cn/lncbook
Contact: lncwiki@big.ac.cn

Related papers and references

LncBook: a curated knowledgebase of human long non-coding RNAs. Nucleic Acids Res 2019.



Data resources of LncBook

11. Information Commons for Rice (IC4R)

Introduction of the data

IC4R is a curated database providing rice genome sequences, updating rice gene annotations and integrating multiple omics data (transcriptome, epigenetics, literature, protein modification) through community-contributed modules.

Format of the data, total size of the data

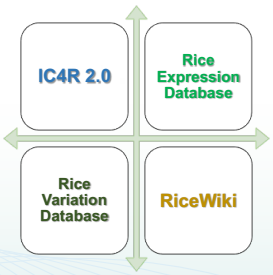
Format: gff, fa, csv
Data size: 1.48 GB

Download site and contact person

http://ic4r.org
Contact: haolili@big.ac.cn

Related papers and references

- IC4R-2.0: rice genome reannotation using massive RNA-seq data. Genomics Proteomics Bioinformatics 2020.
- Information Commons for Rice (IC4R). Nucleic Acids Res 2016.



- **56,221** protein-coding genes
- **80,038** protein-coding transcripts
- **6,259** lncRNAs
- **4,373** circular RNAs
- **1,503** RNA-Seq datasets

Modules and data resources of IC4R

12. iDog

Introduction of the data

iDog is an integrated resource for domestic dog and wild canids, including genes, genomes, SNPs, breed/disease traits, gene expressions, GO function annotations, dog-human homolog diseases and publications. In addition, iDog provides online tools for performing genomic data visualization and analyses.

Format of the data, total size of the data

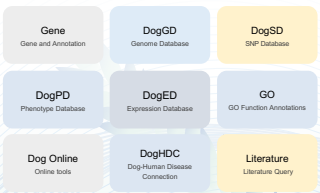
Format: txt, bam, bai, fastq, vcf
Data size: 9.1 TB

Download site and contact person

https://ngdc.cncb.ac.cn/idog
Contact: idog@big.ac.cn

Related papers and references

iDog: an integrated resource for domestic dogs and wild canids. Nucleic Acids Res 2019.



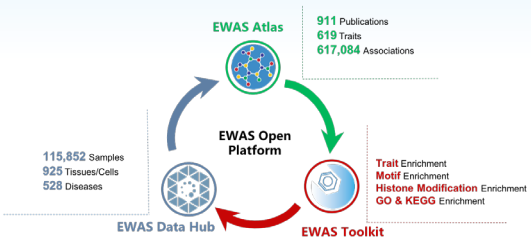
- **481** breeds
- **32,220** genes
- **806** dog diseases
- **1,170** genotype-phenotype pairs
- **7,130** literatures

Modules and data resources of iDog

13. EWAS Open Platform

Introduction of the data

EWAS Open Platform includes three parts: data portal (EWAS Data Hub), knowledge base (EWAS Atlas) and online tools (EWAS Toolkit). EWAS Data Hub is a database for collecting and normalizing DNA methylation array data as well as archiving associated metadata, and provides all terms of DNA methylation profiles for probes and genes. EWAS Atlas is the first manual curated epigenome-wide association study knowledgebase in the world, which contains a large number of phenotypes, environmental factors, behavioral factors and diseases-related EWAS associations and their metadata. Taking advantage of massive high-quality DNA methylation data, EWAS toolkit has been greatly enhanced for a wide range of EWAS analyses.



Overview of EWAS Open Platform

Format of the data, total size of the data

Format: txt
Data size: 1.3 TB

Download site and contact person

https://ngdc.cncb.ac.cn/ewas
Contact: ewas-user@big.ac.cn

Related papers and references

- EWAS Open Platform: integrated data, knowledge and toolkit for epigenome-wide association study. Nucleic Acids Res 2022.
- EWAS Data Hub: a resource of DNA methylation array data and metadata. Nucleic Acids Res 2020.
- EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. Nucleic Acids Res 2019.

14. BrainBase

Introduction of the data

BrainBase is a curated knowledgebase for brain diseases that aims to provide a whole picture of brain diseases and associated genes. It integrates not only valuable curated disease-gene associations and drug-target interactions but also molecular profiles through multi-omics data analysis, accordingly bearing great promise to serve as a valuable knowledgebase for brain diseases.

Format of the data, total size of the data

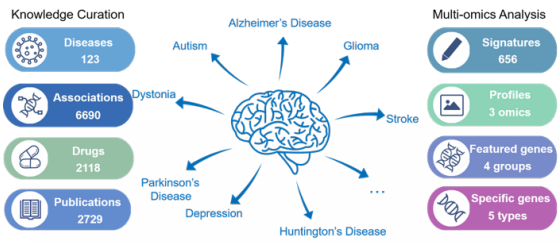
Format: pdf, txt, csv, zip
Data size: 8 GB

Download site and contact person

https://ngdc.cncb.ac.cn/brainbase
Contact: brainbase@big.ac.cn

Related papers and references

BrainBase: a curated knowledgebase for brain diseases. Nucleic Acids Res 2022.



Data resources of BrainBase

15. CancerSCEM

Introduction of the data

CancerSCEM is an open access database of cancer single-cell expression map. Currently CancerSCEM provides comprehensive metadata and multi-scale analyzed results of 208 samples across 20 human cancer types. Equipped with the newly constructed comprehensive online analysis platform, CancerSCEM allows users to perform cancer scRNA-seq data exploration in a real-time and interactive mode.

Format of the data, total size of the data

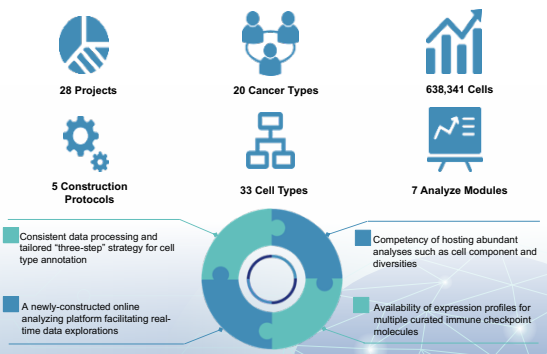
Format: xlsx, jpeg, tsv, zip
Data size: 135 GB

Download site and contact person

https://ngdc.cncb.ac.cn/cancerscem
Contact: zengjy@big.ac.cn

Related papers and references

CancerSCEM: a database of single-cell expression map across various human cancers. Nucleic Acids Res 2022.



Data resources and features of CancerSCEM

16. Single-cell Methylation Bank (scMethBank)

Introduction of the data

scMethBank is a comprehensive database that integrates storage of single-base precision whole-genome methylation data and manually reviewed meta data of single cells. It provides a user-friendly browsing, query, visualization and differential methylation analysis module, which has fundamental significance for medicine, molecular evolution, and comparative epigenomics studies.

Format of the data, total size of the data

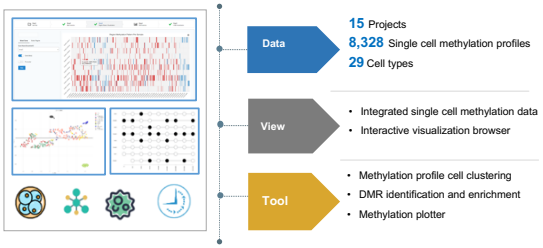
Format: bed.gz
Data size: 3.7 TB

Download site and contact person

https://ngdc.cncb.ac.cn/methbank/scm
Contact: zongwenting2018m@big.ac.cn

Related papers and references

scMethBank: a database for single-cell whole genome DNA methylation maps. Nucleic Acids Res 2022.



Data resources and features of scMethBank

Contact Us

ANSO Secretariat
No. 16 Lincui Road, Chaoyang District, Beijing 100101, China

anso-public@anso.org.cn
http://www.anso.org.cn/

Responsible Editor: Ailikun
Editors: Xin Zhang, Yiming Bao, Ruiyang Zhou, Minyan Zhao
Language Editor: Michael Manton
Issue No. 4 Published in October 2022