

# Biological Databases

生物信息学  
助教-刘柯  
助教-方明昊

# 课程概览

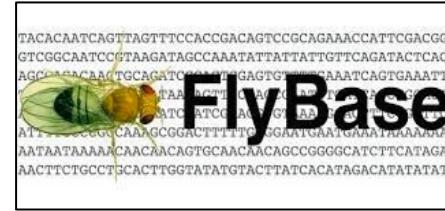
1 数据库基本概念

2 NCBI & EMBL-EBI & 在线分析工具

3 TCGA GTEx & GEPIA 1/2/2021

4 单细胞/空间转录组数据库

# 数据库基本概念



- 数据库 (Database) 是一种结构化的数据存储方式，它允许将大量相关数据**存储、检索、调用和管理**。
- 生物学数据库 (Biological Database) 是一种专门用于**存储和管理生物学数据**的数据库类型。

# 数据库基本概念

JOURNAL ARTICLE

## The 2023 Nucleic Acids Research Database Issue and the online molecular biology database collection

Daniel J Rigden , Xosé M Fernández

*Nucleic Acids Research*, Volume 51, Issue D1, 6 January 2023, Pages D1–D8,

<https://doi.org/10.1093/nar/gkac1186>

**Published:** 06 January 2023

- Nucleic Acids Research每年都会发表数据库特辑，在2023年包括了178篇，包括90篇新的数据库文章，82篇对以前数据库的更新，以及6篇发表在其他杂志的文章。

# 数据库基本概念

The screenshot shows the header "OXFORD ACADEMIC Journals" and the breadcrumb "You are here: NAR Journal Home » Database Summary Paper Categories". Below this is the section title "NAR Database Summary Paper Category List" followed by a list of database categories:

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases
- Cell biology

**每年的特辑发布后，会更新NAR  
数据库网址下收录的各类数据库。**

**在2023年，添加92个新的数据库，  
并且清除96个已经URL停用的数据库。  
目前共收录1764个数据库。**

**有活力的数据库是持续更新和定  
期维护的。**

[http://www.oxfordjournals.org/nar/database  
/c/](http://www.oxfordjournals.org/nar/database/c/)

# 数据库基本概念

## 常见数据库分类

- **序列数据库**

**NCBI-GenBank, DDBJ, EMBL-ENA**(核酸序列数据库)

**UniProt**(蛋白质序列数据库)



- **基因表达数据库**

**GEO**(基因表达综合数据库)

**GTEX**(组织相关表达数据库)

**TCGA**(癌症相关数据库)

**GEPIA/ GEPIA2**(癌症相关数据库)

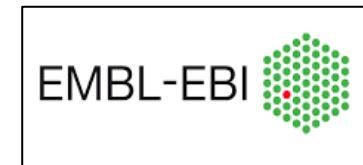
单细胞&空间转录组数据库 (待补充)



- **基因本体数据库GO及注释**

**Gene Ontology Consortium, QuickGO, AmiGo**

GO注释工具: **MetaScape(在线), clusterprofiler, GSEA**



# 数据库基本概念

## 常见数据库分类

- 基因Variation& Mutation数据库  
dbSNP, 1000 Genomes Project, ClinVar
- 代谢及信号转导数据库  
KEGG, Reactome
- 结构数据库  
PDB(蛋白质结构数据库), NDB (核酸结构数据库)
- 蛋白互作数据库  
STRING
- 系统发育数据库  
Tree of Life, NCBI Taxonomy



# 课程概览

1 数据库基本概念

2 NCBI & EMBL-EBI & 在线分析工具

3 TCGA GTEx & GEPIA 1/2/2021

4 单细胞/空间转录组数据库

# NCBI & EBI

---

**NCBI** (National Center for Biotechnology Information) , 隶属于NIH (National Institutes of Health)的NLB (National Library of Medicine), 受到的资助。NCBI成立于1988年，它的主要目标是满足对生物和基因组数据的组织、存储和传播日益增长的需求。

**EMBL-EBI** (European Bioinformatics Institute), 隶属于EMBL (European Molecular Biology Laboratory)。EBI成立于1994年，致力于生物信息学和计算生物学，提供免费、开放的大量生物信息学资源和工具，是生命科学研究人员的重要中心。

NCBI、EMBL-EBI和DDBJ共同运作国际核苷酸序列数据库联盟INSDC，每天都会交换数据。



在1982年，NCBI成立前，GenBank项目启动，并在之后成为NCBI的一部分。1990年代推出Entrez系统，增强了数据检索。在1997年，NCBI推出了PubMed，进入21世纪后，有陆续推出了GEO，SRA等数据库。

**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

**Develop**: Use NCBI APIs and code libraries to build applications

**Analyze**: Identify an NCBI tool for your data analysis task

**Research**: Explore NCBI research and collaborative projects

**Popular Resources**

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

**NCBI News & Blog**

Upcoming Changes to Virus Data Resources at NCBI  
19 Oct 2023  
Effective May 2024, NCBI Virus will replace legacy virus web resources

NCBI Datasets: Easily Access and Download Sequence Data and Metadata  
18 Oct 2023  
Effective May 2024, NCBI Datasets will replace legacy Genome and Assembly

Introducing the New NCBI Datasets Genome Annotation Table  
10 Oct 2023  
As part of our ongoing effort to modernize and improve your experience, we are

[More...](#)

Recent

All Databases

All

- All Databases
- Assembly
- BioCollections
- BioProject
- BioSample
- Books
- ClinVar
- Conserved Domains
- dbGaP
- dbVar
- Gene
- Genome
- GEO DataSets
- GEO Profiles
- GTR
- HomoloGene
- Identical Protein Groups
- MedGen
- MeSH
- NLM Catalog
- Nucleotide
- OMIM
- PMC
- PopSet
- Protein
- Protein Clusters
- Protein Family Models
- PubChem BioAssay
- PubChem Compound
- PubChem Substance
- PubMed
- SNP
- SRA
- Structure
- Taxonomy
- ToolKit

目前一共收录了1000+个物种的整个基因组的数据，包括了生物学研究的主要领域，如细菌、真核生物、病毒等。

The screenshot shows the NCBI Genome homepage. At the top, there is a banner from the National Library of Medicine (NIH) stating: "Effective May 2024, NCBI's Genome resource will no longer be available. NCBI Genome data can now be found on the NCBI Datasets taxonomy pages. [Learn more.](#)" Below this, the main content area has a dark header titled "Genome". A large image of chromosomes is on the left. The central text area says: "This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations." To the right, there is a section titled "Custom resources" with links to "Human Genome", "Microbes", "Organelles", "Viruses", and "Prokaryotic reference genomes". This section is highlighted with a red border. On the far right, there are sections for "Other Resources" (Assembly, BioProject, BioSample, Genome Data Viewer, NCBI Datasets NEW) and "External Resources" (GOLD - Genomes Online Database, Bacteria Genomes at Sanger, Ensembl). Navigation links on the left include "Using Genome" (Help, Browse by Organism UPDATED, Download / FTP, Download FAQ, Submit a genome) and "Genome Tools" (BLAST the Human Genome, Microbial Nucleotide BLAST). The "Genome Annotation and Analysis" section includes links to Eukaryotic Genome Annotation, Prokaryotic Genome Annotation, and PASC (Pairwise Sequence Comparison).

# NCBI-BLAST

Article preview  
Abstract  
References (23)  
Cited by (74453)

**jmb**  
Journal of Molecular Biology

Volume 215, Issue 3, 5 October 1990, Pages 403-410

## Basic local alignment search tool

NIH National Library of Medicine National Center for Biotechnology Information

Log in

BLAST®

Check out the ClusteredNR database on BLAST+ [Learn more](#) [Give us feedback](#)

### 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. [Learn more](#)

NEWS ClusteredNR database on BLAST+ The ClusteredNR database is now available for BLAST+. Thu, 24 Aug 2023 [More BLAST news...](#)

### Web BLAST

**Nucleotide BLAST** nucleotide ▶ nucleotide

**blastx** translated nucleotide ▶ protein

**tblastn** protein ▶ translated nucleotide

**Protein BLAST** protein ▶ protein

### BLAST Genomes

cagggaaaggcagaagaggtcctggctgtgtggggcaggggcaggaaat [Search](#)

Human Mouse Rat Microbes

对于未知的序列，  
NCBI还提供了BLAST  
与NCBI的基因组进行  
比对。

BLAST从1990  
年发表至今，已被引  
用超过7万次。

# NCBI-RefSeq

在GenBank建立的过程中发现不同实验室提交的序列具有冗余的结果，所以建立了RefSeq的数据库。

The screenshot shows two main sections of the NCBI website: the Nucleotide database and the Protein database.

**Nucleotide Database:** The top navigation bar includes the NIH logo, the National Library of Medicine logo, and a "Log in" button. Below the bar, there's a search interface with dropdown menus for "Nucleotide" and "Protein", a search input field, and a "Search" button. A "Help" link is also present. The main content area displays a sequence of DNA bases (A, T, C, G) and is titled "Nucleotide". A descriptive text states: "The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery."

**Protein Database:** Below the Nucleotide section, there's another search interface for "Protein" with similar components: dropdown menus, search input, "Search" button, and "Help" link. The main content area displays a sequence of amino acids (Q, D, L, V, S, R, G, E, I, R, K, T, E, K, T, F, V, P, etc.) and is titled "Protein". A descriptive text states: "The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function."

**Footer Navigation:** At the bottom, there are three columns: "Using Protein" (Quick Start Guide, FAQ, Help, GenBank FTP, RefSeq FTP), "Protein Tools" (BLAST, LinkOut, E-Utilities, Batch Entrez), and "Other Resources" (GenBank Home, RefSeq Home, CDD, Structure).

在RefSeq数据库中能够直接搜索到序列的信息，包括位置信息、物种、功能区域及原始序列。

Nucleotide      Nucleotide  Advanced

GenBank

**Homo sapiens PD1 gene for programmed cell death 1, 3'UTR, partial sequence**

GenBank: LC461712.1  
[FASTA](#) [Graphics](#)

Go to:

LOCUS LC461712 50 bp DNA linear PRI 20-JUL-2019  
DEFINITION Homo sapiens PD1 gene for programmed cell death 1, 3'UTR, partial sequence.  
ACCESSION LC461712  
VERSION LC461712.1  
KEYWORDS .  
SOURCE Homo sapiens (human)  
ORGANISM [Homo sapiens](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.  
REFERENCE 1  
AUTHORS Ahmed,M.Z., Bhardwaj,N., Pande,V., Valecha,N. and Anvikar,A.R.  
TITLE Expression of major transcriptional factors in Plasmodium falciparum infected patients  
JOURNAL Unpublished  
REFERENCE 2 (bases 1 to 50)  
AUTHORS Ahmed,M.Z. and Anvikar,A.R.  
TITLE Direct Submission  
JOURNAL Submitted (03-FEB-2019) Contact: Md Zohaib Ahmed ICMR-National Institute of Malaria Research, Epidemiology and Clinical Research; Sector-8, Dwarka, New Delhi, Delhi 110077, India URL :<http://nimr.org.in/>

FEATURES Location/Qualifiers

source 1..50  
/organism="Homo sapiens"  
/mol\_type="genomic DNA"  
/db\_xref="taxon:9606"  
/chromosome="2"  
/map="2q37.3"  
/country="India"  
/collection\_date="2017"  
/PCR\_primers="fwd\_seq: cctgcaggccatagaaagttt, rev\_seq: ggccatgtttaaagggtgg"  
<1..>50  
/gene="PD1"  
<1..>50  
/gene="PDI"  
/origin="programmed cell death 1"

ORIGIN 1 cagggaaaggc cagaagact cctggctgtg gtgggcaggc caggaaacc

Protein      Protein  Advanced

GenPept

**Chain B, Ribosome-inactivating protein PD-L1/PD-L2**

PDB: 3LE7\_B  
[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to:

LOCUS 3LE7\_B 261 aa linear PLN 01-DEC-2020  
DEFINITION Chain B, Ribosome-inactivating protein PD-L1/PD-L2.  
ACCESSION 3LE7\_B  
VERSION 3LE7\_B  
DBSOURCE pdb: molecule 3LE7, chain B, release Apr 14, 2010; deposition: Jan 14, 2010; class: Hydrolase; source: Mol\_id: 1; Organism\_scientific: Phytolacca Dioica; Organism\_common: Bella Sombra Tree; Organism\_taxid: 29725; Exp. method: X-Ray Diffraction.  
KEYWORDS .  
SOURCE Phytolacca dioica  
ORGANISM [Phytolacca dioica](#)  
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; Caryophyllales; Phytolaccaceae; Phytolacca.  
REFERENCE 1 (residues 1 to 261)  
AUTHORS Severino,V., Chambery,A., Di Maro,A., Marasco,D., Ruggiero,A., Berisio,R., Giansanti,F., Ippoliti,R. and Parente,A.  
TITLE The role of the glycan moiety on the structure-function relationships of PD-L1, type 1 ribosome-inactivating protein from P. dioica leaves  
JOURNAL Mol Biosyst 6 (3), 570-579 (2010)  
PUBMED 20174685  
REFERENCE 2 (residues 1 to 261)  
AUTHORS Ruggiero,A. and Berisio,R.  
TITLE Direct Submission  
JOURNAL Submitted (14-JAN-2010)  
COMMENT Crystal structure of PD-L1 from P. dioica in complex with adenine.  
FEATURES Location/Qualifiers

source 1..261  
/organism="Phytolacca dioica"  
/db\_xref="taxon:29725"  
SecStr 2..8  
/sec\_str\_type="sheet"  
/note="strand 1"  
Het bond(10)  
/heterogen="(NAG,1003)"  
Region 12..214  
/region\_name="RIP"  
/note="Ribosome inactivating protein; pfam00161"

在NCBI-Gene数据库中，包括了Archaea、Bateria、Eukaryota、Viruses等的基因数目超过5000万个。

 National Library of Medicine  
National Center for Biotechnology Information

Gene



**Gene**

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

**Using Gene**

- [Gene Quick Start](#)
- [FAQ](#)
- [Download/FTP](#)
- [RefSeq Mailing List](#)
- [Gene News](#) 
- [Factsheet](#)

**Gene Tools**

- [Submit GeneRIFs](#)
- [Submit Correction](#)
- [Statistics](#)
- [BLAST](#)
- [Genome Workbench](#)
- [Splign](#)

**Other Resources**

- [OMIM](#)
- [RefSeq](#)
- [RefSeqGene](#)
- [Protein Clusters](#)

Taxa that satisfy your query:

Taxa	Count of Child Taxa
Archaea	504
Bacteria	1813
Eukaryota	30885
Viruses	14382

Total Genes of all children
1646475
5684429
40612523
691667

在NCBI-Gene数据中，可以知道基因的种系、基本功能、染色体的位置及组织表达等。

The screenshot shows the NCBI Gene search interface. At the top, there's a header with the NIH National Library of Medicine logo and the text "National Center for Biotechnology Information". Below the header, a search bar has "Gene" selected and contains the query "PDCD1". There are links for "Full Report" and "Send to:".

The main content area displays information for the gene PDCD1. It includes:

- Official Symbol:** PDCD1 provided by HGNC
- Official Full Name:** programmed cell death 1 provided by HGNC
- Primary source:** HGNC;HGNC:8760
- See related:** Ensembl:ENSG00000188389 MIM:600244; AllianceGenome:HGNC:8760
- Gene type:** protein coding
- RefSeq status:** REVIEWED
- Organism:** Homo sapiens
- Lineage:** Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
- Also known as:** PD1; PD-1; CD279; SLEB2; hPD-1; hPD-L1; hSLE1
- Summary:** Programmed cell death protein 1 (PDCD1) is an immune-inhibitory receptor expressed in activated T cells; it is involved in the regulation of T-cell functions, including those of effector CD8+ T cells. In addition, this protein can also promote the differentiation of CD4+ T cells into T regulatory cells. PDCD1 is expressed in many types of tumors including melanomas, and has demonstrated to play a role in anti-tumor immunity. Moreover, this protein has been shown to be involved in safeguarding against autoimmunity; however, it can also contribute to the inhibition of effective anti-tumor and anti-microbial immunity. [provided by RefSeq, Aug 2020]
- Expression:** Biased expression in lymph node (RPKM 8.8), spleen (RPKM 2.6) and 7 other tissues. [See more](#)
- Orthologs:** mouse, all

A "NEW" button with the text "Try the new Gene table" is present. Below the summary, there's a "Genomic context" section showing the location (2q37.3), exon count (6), and a table of annotation releases. The table has columns for Annotation release, Status, Assembly, Chr, and Location. Rows include RS\_2023\_10 (current, GRCh38.p14, chr2, NC\_000002.12), RS\_2023\_10 (current, T2T-CHM13v2.0, chr2, NC\_060926.1), and 105.20220307 (previous assembly, GRCh37.p13, chr2, NC\_000002.11).

# GeneCards

除NCBI-Gene外，如果想要查询人的基因的信息，也常使用GeneCards数据库。

Free for academic non-profit institutions. Other users need a [Commercial license](#)

WEIZMANN INSTITUTE OF SCIENCE  LifeMap SCIENCES 

Search GeneCards (supports boolean, parenthesis and quotes)  Advanced

Home | Analysis Tools ▾ | Release Notes | About ▾ | Data Access | GeneCards Team | Help ▾ | My Genes | Log In / Sign Up

## PDCD1 Gene - Programmed Cell Death 1

Protein Coding (Updated: Oct 4, 2023 ; GC02M241849 ⓘ ; GIfTS: 51 ⓘ)  

Jump to section: Aliases, Disorders, Domains, Drugs, Expression, Function Sources, Genomics Summaries, Localization Transcripts, Orthologs Variants.

Research Products: Antibodies, Assays, Proteins, Inhib. RNA, CRISPR, miRNA, Drugs, Animal Models, Cell Lines.

R&D: Proteins Primary Antibodies ELISAs, Antibody Arrays Activity Assays.

VectorBuilder: Online Vector Design Platform, Virus Packaging (AAV/Lenti), CRISPR Library Construction.

ORIGENE: Proteins Antibodies Assays Genes, shRNA Primers CRISPR, Lentiviral Particles.

SYNTHEGO: CRISPR Knockout Kit sgRNA, Engineered Cells Edited iPSCs, Free Bioinformatics Tools.

### Aliases for PDCD1 Gene

Aliases for PDCD1 Gene

GeneCards Symbol: **PDCD1** ⓘ

Programmed Cell Death 1 ⓘ

PD1 ⓘ ⓘ ⓘ ⓘ

CD279 ⓘ ⓘ ⓘ

HSLE1 ⓘ ⓘ ⓘ

PD-1 ⓘ ⓘ ⓘ

Systemic Lupus Erythematosus Susceptibility 2 ⓘ ⓘ

External IDs for PDCD1 Gene

HGNC: 8760 NCBI Gene: 5133 Ensembl: ENSG00000188389 OMIM: 600244 UniProtKB/Swiss-Prot: Q15116

Previous HGNC Symbols for PDCD1 Gene

SLEB2

Previous GeneCards Identifiers for PDCD1 Gene

GC02U990074, GC02P241575, GC02P9D0113, GC02M242440, GC02M242792, GC02M234541

Search aliases for PDCD1 gene in PubMed and other databases

### GeneCards for Bioinformaticians

Annotate your datasets with comprehensive integrated data from >190 biomedical sources

GENES  
DISEASES  
PATHWAYS  
ENHancers  
VARIANTS

GET STARTED > 

在阅读文献的过程中，研究者也会将项目产生的数据上传到**GEO**的数据库。  
目前国内的研究者会将数据上传到**GSA**数据库。

Sections      Figures      References

Abstract

Main

Technology workflow and data quality

Spatial comapping of mouse embryo

Spatial ATAC-RNA-seq of mouse brain

Spatial CUT&Tag-RNA-seq of mouse brain

Region-specific gene expression regulation

Spatial comapping of human brain

Discussion

Methods

**Data availability**

Code availability

## Data availability

Raw and processed data reported in this paper are deposited in the Gene Expression Omnibus with accession code [GSE205055](#). These datasets are available as web resources and can be browsed within the tissue spatial coordinates in the UCSC Cell and Genome Browser (<https://brain-spatial-omics.cells.ucsc.edu>), and in our own data portal generated with AtlasXplore (<https://web.atlasxomics.com/visualization/Fan>). Data are also available at <https://ki.se/en/mbb/oligointernode>. The resulting fastq files were aligned to either the human reference genome (GRCh38) (<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/>) or mouse reference genome (GRCm38) (<https://hgdownload.soe.ucsc.edu/goldenPath/mm10/chromosomes/>).

Published data for integration and quality comparison are available online: ENCODE ATAC-seq (E13.5 mouse embryo) (<https://www.encodeproject.org/search/?>

Zhang D, Deng Y, Kukanja P, et al. *Nature*, 2023.

在NCBI输入文章提供的GSE号，就能够搜索到对应的数据集。

 National Library of Medicine  
National Center for Biotechnology Information

GEO DataSets      GEO DataSets ▾ GSE205055 | Create alert Advanced

Entry type      Summary ▾ 20 per page ▾ Sort by Default order ▾      Send to: ▾  
DataSets (0)  
Series (1)  
Samples (22)  
Platforms (2)

Organism      **Search results**  
Customize ...

Study type      Items: 1 to 20 of 25      << First < Prev Page  of 2 Next > Last >>  
Expression profiling by array  
Methylation profiling by array  
Customize ...

[Spatial epigenome-transcriptome co-profiling of mammalian tissues](#)  
1. (Submitter supplied) This SuperSeries is composed of the SubSeries listed below.  
Organism: Homo sapiens; Mus musculus  
Type: Expression profiling by high throughput sequencing; Genome binding/occupancy profiling by high throughput sequencing  
Platforms: GPL24676 GPL24247 22 Samples  
[Download data: TAR, TSV](#)

Author      Series Accession: **GSE205055** ID: 200205055  
Customize ...

Attribute name      [PubMed](#) [Full text in PMC](#) [Similar studies](#)

在对应的GEO数据库中，能够知道对应的文章、物种、实验类型、测序平台、实验样本等信息。在supplementary file一般会存储处理后的矩阵数据。SRA Run selector可以查看上传到NCBI-SRA数据库的原始数据。

[HOME](#) | [SEARCH](#) | [SITE MAP](#)

NCBI > GEO > **Accession Display** ⓘ

GEO help: Mouse over screen elements for information.

Scope: **Self** Format: **HTML** Amount: **Quick** GEO accession: **GSE205055** Go

**Series GSE205055** Query DataSets for GSE205055

Status	Public on Jan 29, 2023		
Title	Spatial epigenome-transcriptome co-profiling of mammalian tissues		
Organisms	Homo sapiens; Mus musculus		
Experiment type	Expression profiling by high throughput sequencing Genome binding/occupancy profiling by high throughput sequencing		
Summary	This SuperSeries is composed of the SubSeries listed below.		
Overall design	Refer to individual Series		
Citation(s)	Zhang D, Deng Y, Kukanja P, Agirre E et al. Spatial epigenome-transcriptome co-profiling of mammalian tissues. <i>Nature</i> 2023 Apr;616(7955):113-122. PMID: <a href="#">36922587</a>		
Submission date	May 29, 2022		
Last update date	Apr 14, 2023		
Contact name	Di Zhang		
E-mail(s)	<a href="mailto:di.zhang@yale.edu">di.zhang@yale.edu</a>		
Organization name	Yale University		
Department	Biomedical Engineering		
Street address	55 Prospect St		
City	New Haven		
State/province	CT		
ZIP/Postal code	06511		
Country	USA		
Platforms (2)	<a href="#">GPL24247</a> Illumina NovaSeq 6000 (Mus musculus) <a href="#">GPL24676</a> Illumina NovaSeq 6000 (Homo sapiens)		
Samples (22)	<a href="#">GSM6204621</a> MouseBrain_20um_H3K27ac <a href="#">GSM6204623</a> MouseBrain_20um <a href="#">GSM6204624</a> ME13_100barcodes_25um		
This SuperSeries is composed of the following SubSeries:			
<a href="#">More...</a>			
<a href="#">GSE205051</a> Spatial epigenome-transcriptome co-profiling of mammalian tissues [CUT&TAG]			
<a href="#">GSE205052</a> Spatial epigenome-transcriptome co-profiling of mammalian tissues [ATAC-Seq]			
<a href="#">GSE205054</a> Spatial epigenome-transcriptome co-profiling of mammalian tissues [RNA-Seq]			
Relations	<a href="#">PRJNA843455</a>		
Download family	Format		
SOF formatted family file(s)	SOFT ⓘ		
MINIMI formatted family file(s)	MINIMI ⓘ		
Series Matrix File(s)	TXT ⓘ		
<b>Supplementary file</b>	<b>Size</b>	<b>Download</b>	<b>File type/resource</b>
GSE205055_RAW.tar	7.6 Gb	(http)(custom)	TAR (of TAR, TSV)
<a href="#">SRA Run Selector</a> ⓘ			

Supplementary file	Size	Download	File type/resource
GSE205055_RAW.tar	7.6 Gb	(http)(custom)	TAR (of TAR, TSV)
<a href="#">SRA Run Selector</a> ⓘ			
Custom GSE205055_RAW.tar archive:			
Supplementary file			File size
<input type="checkbox"/> <a href="#">GSM6204621_MouseBrain_20um_H3K27ac_fragments.tsv.gz</a>			147.8 Mb
<input type="checkbox"/> <a href="#">GSM6204621_MouseBrain_20um_H3K27ac_spatial.tar.gz</a>			20.6 Mb
<input type="checkbox"/> <a href="#">GSM6204623_MouseBrain_20um_fragments.tsv.gz</a>			329.0 Mb
<input type="checkbox"/> <a href="#">GSM6204623_MouseBrain_20um_spatial.tar.gz</a>			17.7 Mb
<input type="checkbox"/> <a href="#">GSM6204624_ME13_100barcodes_25um_fragments.tsv.gz</a>			1.5 Gb
<input type="checkbox"/> <a href="#">GSM6204624_ME13_100barcodes_25um_spatial.tar.gz</a>			25.9 Mb
<input type="checkbox"/> <a href="#">GSM6204635_MouseBrain_20um_H3K27ac_matrix.tsv.gz</a>			4.0 Mb
<input type="checkbox"/> <a href="#">GSM6204635_MouseBrain_20um_H3K27ac_spatial.tar.gz</a>			20.6 Mb
<input type="checkbox"/> <a href="#">GSM6204636_MouseBrain_20um_matrix.tsv.gz</a>			3.6 Mb
<input type="checkbox"/> <a href="#">GSM6204636_MouseBrain_20um_spatial.tar.gz</a>			17.7 Mb
<input type="checkbox"/> <a href="#">GSM6204637_ME13_100barcodes_25um_matrix.tsv.gz</a>			7.9 Mb
<input type="checkbox"/> <a href="#">GSM6204637_ME13_100barcodes_25um_spatial.tar.gz</a>			25.9 Mb
<input type="checkbox"/> <a href="#">GSM6206884_HumanBrain_50um_fragments.tsv.gz</a>			435.3 Mb
<input type="checkbox"/> <a href="#">GSM6206884_HumanBrain_50um_spatial.tar.gz</a>			49.2 Mb
<input type="checkbox"/> <a href="#">GSM6206885_HumanBrain_50um_matrix.tsv.gz</a>			4.9 Mb
<input type="checkbox"/> <a href="#">GSM6206885_HumanBrain_50um_spatial.tar.gz</a>			49.2 Mb
<input type="checkbox"/> <a href="#">GSM6704977_MouseBrain_20um_rep_H3K27ac_fragments.tsv.gz</a>			158.2 Mb
<input type="checkbox"/> <a href="#">GSM6704978_MouseBrain_20um_100barcodes_H3K27me3_fragments.tsv.gz</a>			1.1 Gb
<input type="checkbox"/> <a href="#">GSM6704979_MouseBrain_20um_100barcodes_H3K27ac_fragments.tsv.gz</a>			1.0 Gb
<input type="checkbox"/> <a href="#">GSM6704980_MouseBrain_20um_100barcodes_H3K27me3_fragments.tsv.gz</a>			262.6 Mb
<input type="checkbox"/> <a href="#">Select All</a>			
	<a href="#">Cancel</a>	<a href="#">Download</a>	
<b>0 file(s), 0 b</b>			

测序的原始数据一般存储在NCBI-SRA数据库中，可以查看到实验建库方式、以及的原始测序文件大小（碱基数/字节数）。

**Select**

	Runs	Bytes	Bases	Download
Total	20	824.04 Gb	2.41 T	Metadata or Accession List
Selected	0	0	0	Metadata or Accession List or JWT C

Found 20 Items

Run	BioProject	BioSample	Assay Type	Bases	Bytes
SRR19441271	PRJNA843459	SAMN28742910	OTHER	39.22 G	11.57 Gb
SRR19441281	PRJNA843456	SAMN28742906	RNA-Seq	61.18 G	21.07 Gb
SRR19441282	PRJNA843456	SAMN28742907	RNA-Seq	34.53 G	11.83 Gb
SRR19441283	PRJNA843456	SAMN28742908	RNA-Seq	31.45 G	10.84 Gb
SRR19441285	PRJNA843458	SAMN28742911	ATAC-seq	82.20 G	25.46 Gb
SRR19441286	PRJNA843458	SAMN28742912	ATAC-seq	38.20 G	12.12 Gb
SRR22141530	PRJNA896927	SAMN31572405	OTHER	195.81 G	62.57 Gb
SRR22141531	PRJNA896927	SAMN31572406	OTHER	139.73 G	42.41 Gb
SRR22141532	PRJNA896927	SAMN31572407	OTHER	152.43 G	49.11 Gb
SRR22141533	PRJNA896927	SAMN31572408	OTHER	49.66 G	15.79 Gb
SRR22385469	PRJNA904377	SAMN31841204	RNA-Seq	280.77 G	100.22 Gb
SRR22385470	PRJNA904377	SAMN31841205	RNA-Seq	311.59 G	110.71 Gb
SRR22385471	PRJNA904377	SAMN31841206	RNA-Seq	302.13 G	108.74 Gb
SRR22385472	PRJNA904377	SAMN31841207	RNA-Seq	357.01 G	129.36 Gb
SRR22385473	PRJNA904377	SAMN31841208	RNA-Seq	42.65 G	14.91 Gb
SRR22385474	PRJNA904377	SAMN31841209	RNA-Seq	40.27 G	14.08 Gb
SRR22428572	PRJNA843458	SAMN31891896	ATAC-seq	121.63 G	39.29 Gb
SRR22428573	PRJNA843458	SAMN31891897	ATAC-seq	49.55 G	15.88 Gb

National Library of Medicine  
National Center for Biotechnology Information

SRA SRA SRR19441271 Create alert Advanced

Full ▾ Send to: ▾

**SRX15494451: GSM6204621: MouseBrain\_20um\_H3K27ac; Mus musculus; OTHER**  
1 ILLUMINA (Illumina NovaSeq 6000) run: 130.7M spots, 39.2G bases, 11.6Gb downloads

**External Id:** GSM6204621\_r1  
**Submitted by:** Biomedical Engineering, Yale University  
**Study:** Spatial epigenome-transcriptome co-profiling of mammalian tissues [CUT&TAG]  
[PRJNA843459](#) • [SRP377553](#) • [All experiments](#) • [All runs](#)  
[show Abstract](#)

**Sample:** MouseBrain\_20um\_H3K27ac  
[SAMN28742910](#) • [SRS13209545](#) • [All experiments](#) • [All runs](#)  
**Organism:** *Mus musculus*

**Library:**  
**Name:** GSM6204621  
**Instrument:** Illumina NovaSeq 6000  
**Strategy:** OTHER  
**Source:** GENOMIC  
**Selection:** other  
**Layout:** PAIRED

**Construction protocol:** The fresh frozen tissue section was fixed with formaldehyde. The adapters loaded pA-Tn5 transposition containing a galvan linker was inserted into transposase accessible genomic DNA loci. The biotin decorated poly-T primer with a ligation linker was added to capture the mRNA for reverse transcription. In situ ligation was conducted by introducing distinct spatial barcode Ai (i = 1-50/100) to the adaptors through an array of lateral microchannels. Then, distinct spatial barcode Bj (j = 1-50/100) were introduced through the longitudinal microchannels. Barcodes A and B with linkers were ligated to the 5' end of the pA-Tn5 oligo separately during each ligation. The tissue can be spatially barcoded with a distinct combination of barcodes Ai and Bj (i = 1-50/100, j = 1-50/100, n of barcoded pixels = 2,500/10,000). To correlate the spatially barcoded accessibility, transcriptome, and tissue morphology, the tissue was imaged after each ligation. Reverse crosslinking was then performed to collect barcoded cDNA and DNA fragments. The streptavidin beads were used to separate DNA and cDNA fragments. The cDNA fragments can be enriched with streptavidin beads and the DNA fragment was left in the supernatant. After separation, the DNA and cDNA libraries were constructed separately during PCR amplification. NGS sequencing was then performed using a NovaSeq 6000 sequencer with pair-end 150 bp mode.

Run : 1 run, 130.7M spots, 39.2G bases, 11.6Gb

Run	# of Spots	# of Bases	Size	Published
SRR19441271	130,727,338	39.2G	11.6Gb	2023-01-30

# NCBI-SRA

对于数量较少、内存较小的数据，可以直接在SRA Run Selector中选择对应的SRR号，直接在网页进行下载。

National Library of Medicine  
National Center for Biotechnology Information

Sequence Read Archive    Search    Run Browser    Analyses    Study    Provisional SRA    Documentation    Mirroring

Run Browser > SRR19441271

### GSM6204621: MouseBrain\_20um\_H3K27ac; Mus musculus; OTHER (SRR19441271)

Metadata    Reads    Data access    **FASTA/FASTQ download**

Run

Run	Spots	Bases	Size	GC Content	Published	Access Type
SRR19441271	130.7M	39.2G	11.6G	52.2%	2023-01-30	public

Quality graph (bigger)

This run has 2 reads per spot:

L=150, 100%    L=150, 100%

Legend

# NCBI-SRA

对于数量较多、内存较大的数据，可以直接在NCBI-SRA数据库中下载SRA Toolkit，使用它下载SRA数据库中的原始数据。SRA-Toolkit的安装及使用在Github具有详细教程。<https://github.com/ncbi/sra-tools>

The screenshot shows the NCBI SRA homepage. At the top, there's a blue header with the NIH National Library of Medicine logo and the text "National Center for Biotechnology Information". On the right side of the header is a "Log in" button. Below the header is a search bar with the text "SRA" in it, a dropdown menu set to "SRA", and a "Search" button. To the right of the search bar is a "Help" link. The main content area features a large image of a glowing blue DNA helix. To the right of the image, the text "SRA - Now available on the cloud" is displayed, followed by a detailed description of the Sequence Read Archive (SRA) data. Below this section are three columns of links: "Getting Started", "Tools and Software", and "Related Resources". The "Tools and Software" column contains links for "Download SRA Toolkit", "SRA Toolkit Documentation" (which is highlighted with a red border), "SRA-BLAST", "SRA Run Browser", and "SRA Run Selector".

**Getting Started**

- [Documentation](#)
- [How to submit](#)
- [How to search and download](#)
- [How to use SRA in the cloud](#)
- [Submit to SRA](#)

**Tools and Software**

- [Download SRA Toolkit](#)
- [SRA Toolkit Documentation](#)
- [SRA-BLAST](#)
- [SRA Run Browser](#)
- [SRA Run Selector](#)

**Related Resources**

- [Submission Portal](#)
- [dbGaP Home](#)
- [BioProject](#)
- [BioSample](#)

安装好SRA-Toolkit后，在SRA Run Selector中选择需要下载的数据，导出数据对应的SRR号，使用SAR Toolkit的prefetch命令和fasterq dump下载原始的SRA数据并将sra格式的数据转换为fastq格式。

The screenshot shows the NCBI SRA Run Selector interface. At the top, there is a search bar with the placeholder "Search SRA Data" and a "Run ID" input field. Below the search bar are two tabs: "Run ID" and "Accession ID". The "Run ID" tab is selected. The main area displays a table titled "Found 20 Items" with the following columns: Run ID, BioProject, BioSample, Assay Type, Bases, Bytes, Center Name, Experiment, Library Name, and Library Selection. The first 14 rows of the table are highlighted with a red border. The last six rows are standard white rows.

Run ID	BioProject	BioSample	Assay Type	Bases	Bytes	Center Name	Experiment	Library Name	Library Selection
SRR19441271	PRJNA843459	SAMN28742910	OTHER	39.22 G	11.57 Gb	BIOMEDICAL ENGINEERING, YALE UNIVERSITY	SRX15494451	GSM6204621	other
SRR19441281	PRJNA843456	SAMN28742906	RNA-Seq	61.18 G	21.07 Gb	BIOMEDICAL ENGINEERING, YALE UNIVERSITY	SRX15494464	GSM6204637	cDNA
SRR19441282	PRJNA843456	SAMN28742907	RNA-Seq	34.53 G	11.83 Gb	BIOMEDICAL ENGINEERING, YALE UNIVERSITY	SRX15494463	GSM6204636	cDNA
SRR19441283	PRJNA843456	SAMN28742908	RNA-Seq	31.45 G	10.84 Gb	BIOMEDICAL ENGINEERING, YALE UNIVERSITY	SRX15494462	GSM6204635	cDNA
SRR19441285	PRJNA843458	SAMN28742911	ATAC-seq	82.20 G	25.46 Gb	BIOMEDICAL ENGINEERING, YALE UNIVERSITY	SRX15494467	GSM6204624	other
SRR19441286	PRJNA843458	SAMN28742912	ATAC-seq	38.20 G	12.12 Gb	BIOMEDICAL ENGINEERING, YALE UNIVERSITY	SRX15494466	GSM6204623	other
SRR22141530	PRJNA896927	SAMN31572405	OTHER	195.81 G	62.57 Gb	YALE UNIVERSITY	SRX18120858	GSM6704980	other
SRR22141531	PRJNA896927	SAMN31572406	OTHER	139.73 G	42.41 Gb	YALE UNIVERSITY	SRX18120857	GSM6704979	other
SRR22141532	PRJNA896927	SAMN31572407	OTHER	152.43 G	49.11 Gb	YALE UNIVERSITY	SRX18120856	GSM6704978	other
SRR22141533	PRJNA896927	SAMN31572408	OTHER	49.66 G	15.79 Gb	YALE UNIVERSITY	SRX18120855	GSM6704977	other
SRR22385469	PRJNA904377	SAMN31841204	RNA-Seq	280.77 G	100.22 Gb	BIOMEDICAL ENGINEERING, YALE UNIVERSITY	SRX18355555	GSM6753046	cDNA
SRR22385470	PRJNA904377	SAMN31841205	RNA-Seq	311.59 G	110.71 Gb	BIOMEDICAL ENGINEERING, YALE UNIVERSITY	SRX18355554	GSM6753045	cDNA
SRR22385471	PRJNA904377	SAMN31841206	RNA-Seq	302.13 G	108.74 Gb	BIOMEDICAL ENGINEERING, YALE UNIVERSITY	SRX18355553	GSM6753044	cDNA
SRR22385472	PRJNA904377	SAMN31841207	RNA-Seq	357.01 G	129.36 Gb	BIOMEDICAL ENGINEERING, YALE UNIVERSITY	SRX18355552	GSM6753043	cDNA

# NCBI-SRA

或者在GEO数据库中搜索GSE号，勾选对应的样本，导出summary file。找到对应的ftp下载地址，使用wget或者curl命令下载。

National Library of Medicine  
National Center for Biotechnology Information

GEO DataSets | GSE205055 | Search | Create alert | Advanced

Entry type: Summary | 20 per page | Sort by Default order

Search results: Items: 1 to 20 of 25 Selected: 2

1. Spatial epigenome-transcriptome co-profiling of mammalian tissues (Submitter supplied) This SuperSeries is composed of the SubSeries listed below.

Organism: Homo sapiens; Mus musculus  
Type: Expression profiling by high throughput sequencing; Genome binding/occupancy profiling by high throughput sequencing  
Platform: GPL24676 GPL24247 22 Samples  
Download data: TAR, TSV  
Series Accession: GSE205055 ID: 200205055  
Platform Accession: GPL24676 ID: 100024676

2. Illumina NovaSeq 6000 (Homo sapiens)  
Organism: Homo sapiens  
11242 Series 756183 Samples  
Download data  
Platform Accession: GPL24676 ID: 100024676

3. Illumina NovaSeq 6000 (Mus musculus)  
Organism: Mus musculus  
8950 Series 589482 Samples  
Download data  
Platform Accession: GPL24247 ID: 100024247

4. RNA-Seq HumanBrain\_50um  
Organism: Homo sapiens  
Source name: Human hippocampus  
Platform: GPL24676 Series: GSE205055 GSE205181  
Download data: TAR, TSV  
Sample Accession: GSM6206885 ID: 306206885  
SRA Run Selector

5. ATAC-Seq HumanBrain\_50um  
Organism: Homo sapiens  
Source name: Human hippocampus  
Platform: GPL24676 Series: GSE205055 GSE205180  
Download data: TAR, TSV

```
1. Spatial epigenome-transcriptome co-profiling of mammalian tissues
(Submitter supplied) This SuperSeries is composed of the SubSeries listed below.
Organism: Homo sapiens; Mus musculus
Type: Expression profiling by high throughput sequencing; Genome binding/
occupancy profiling by high throughput sequencing
Platform: GPL24676 GPL24247 22 Samples
FTP download: GEO (TAR, TSV) ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE205nnn/
GSE205055

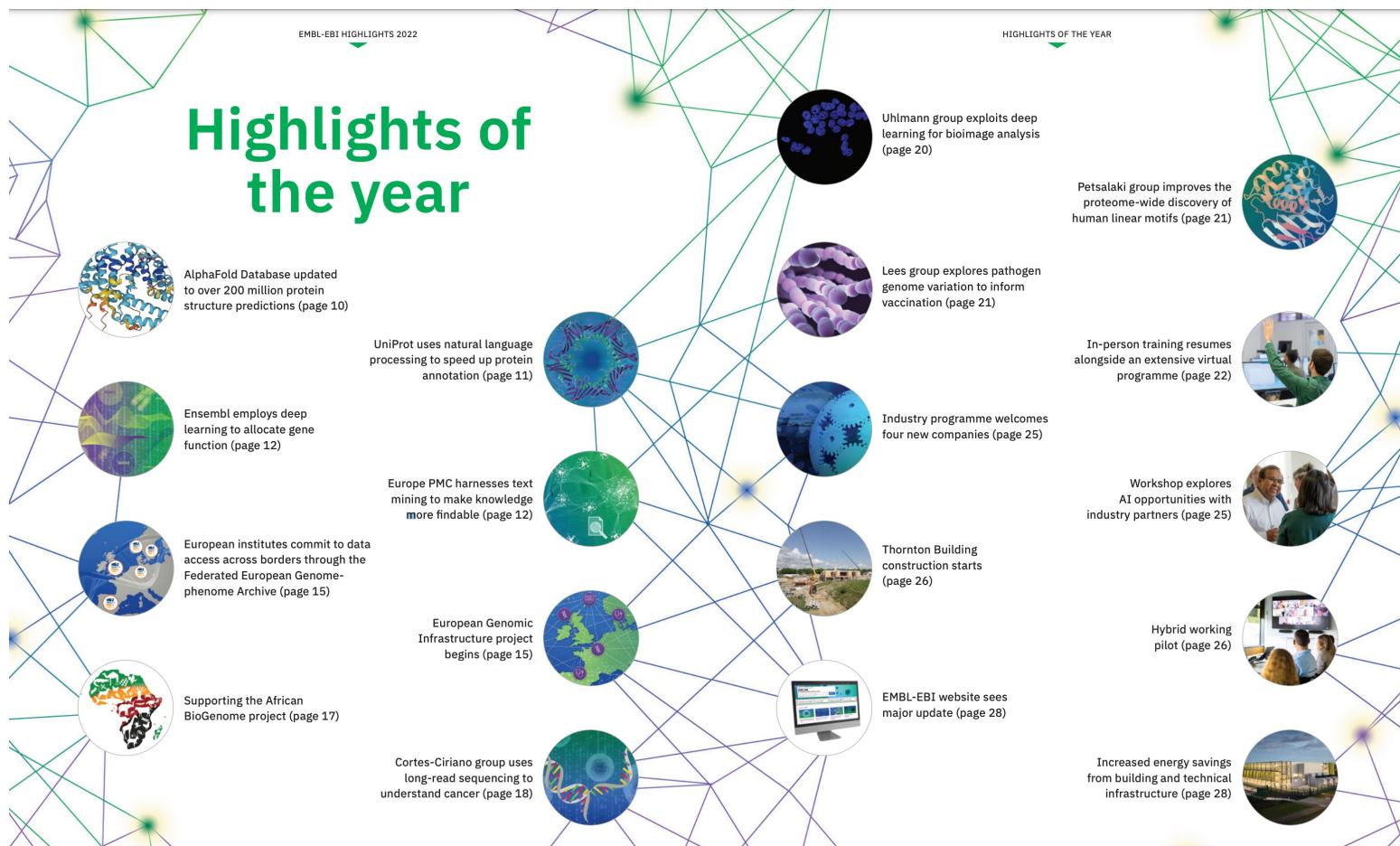
2. Illumina NovaSeq 6000 (Homo sapiens)
Organism: Homo sapiens
11242 Series 756183 Samples
FTP download: GEO ftp://ftp.ncbi.nlm.nih.gov/geo/platforms/GPL24nnn/GPL24676/
Platform Accession: GPL24676 ID: 100024676

3. Illumina NovaSeq 6000 (Mus musculus)
Organism: Mus musculus
8950 Series 589482 Samples
FTP download: GEO ftp://ftp.ncbi.nlm.nih.gov/geo/platforms/GPL24nnn/GPL24247/
Platform Accession: GPL24247 ID: 100024247

4. RNA-Seq HumanBrain_50um
Organism: Homo sapiens
Source name: Human hippocampus
Platform: GPL24676 Series: GSE205055 GSE205181
FTP download: GEO (TAR, TSV) ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM6206nnn/
GSM6206885/
SRA Run Selector: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX15507575
Sample Accession: GSM6206885 ID: 306206885

5. ATAC-Seq HumanBrain_50um
Organism: Homo sapiens
```

EBI位于伦敦剑桥，与Sanger Institute共享区域，共同收到Wellcome Trust的资金资助。在2022年中highlight成果中包括了AlphaFold数据库，Ensembl使用深度学习分配基因功能、Uniprot数据库使用自然语言处理加入蛋白注释等。



与NCBI相似，EBI除了原始的ENA数据库外，还有AlphaFold DB、UniProt等蛋白质相关的数据库。同时也有ArrayExpress、EGA等基因表达相关的数据库。以及Reactome细胞信号通路相关的数据库还有GO相关的quickGO数据库。

## Explore all our data resources and tools →

Our full range of data resources and data analysis tools are essential for supporting life science research.

### Featured data resources



#### AlphaFold DB

Database for protein structure predictions for numerous species

CC-BY



#### ArrayExpress

A database of functional genomics experiments, including microarray and RNAseq expression data typically related to publications.

Web API



#### BioImage Archive

The BioImage Archive is EMBL-EBI's general purpose image archive, accepting molecular biology imaging data associated with peer-reviewed publications.

### Featured tools



#### Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

Web API



#### HMMER

Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.

Web API



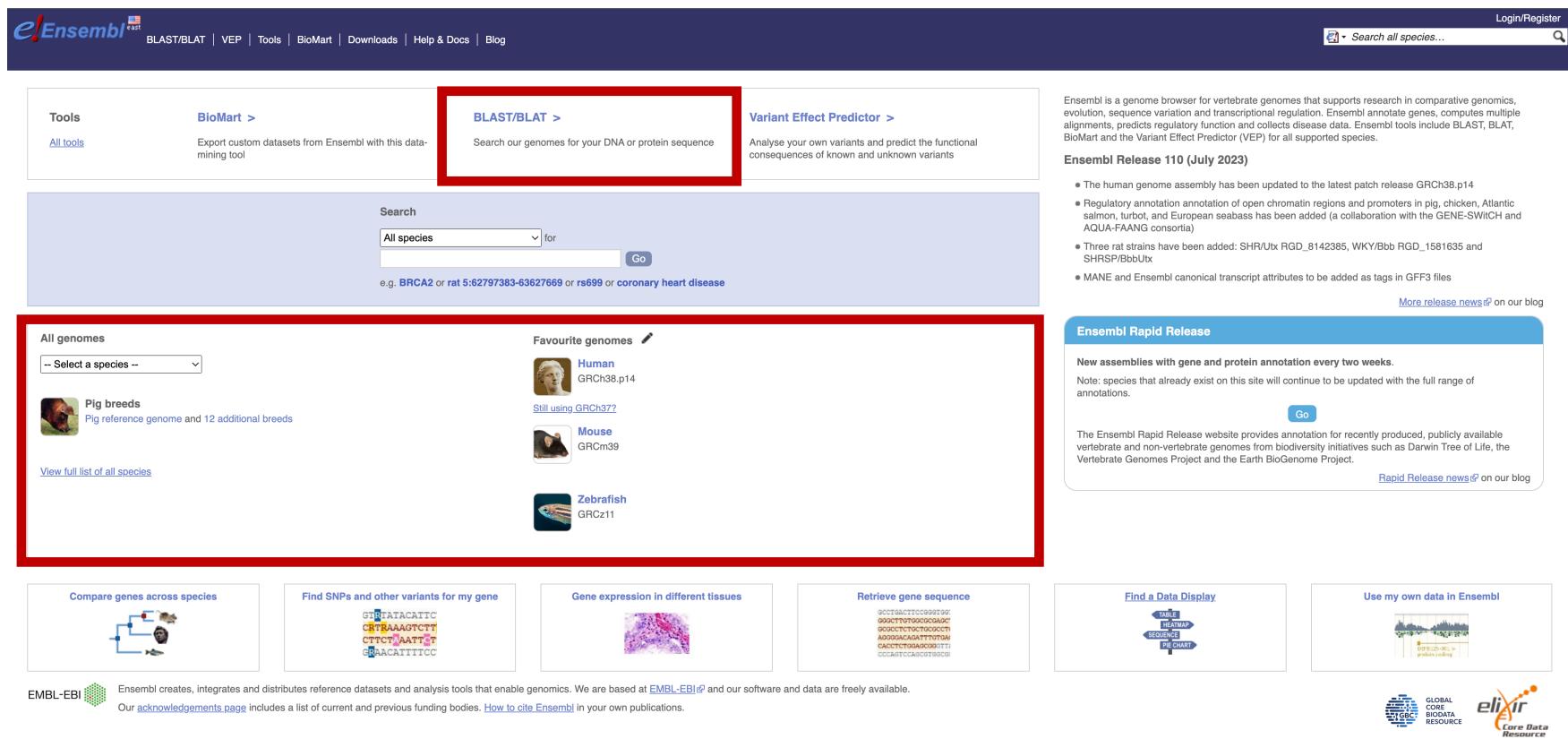
#### Annotation Platform

Consolidating text-mined and curated annotations

Web API

# EMBL-EBI Ensembl

与NCBI-Genome类似，在Ensembl数据库中，存储了大量物种的基因组信息。可以直接在网页选择对应的基因组进行下载。同样在Ensembl的网页也提供BLAST工具。



The screenshot shows the Ensembl homepage. At the top, there is a navigation bar with links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. On the right side of the header is a search bar labeled "Search all species..." and a "Login/Register" button.

The main content area has several sections:

- Tools**: A link to "All tools".
- BioMart >**: A link to "Export custom datasets from Ensembl with this data-mining tool".
- BLAST/BLAT >**: A link to "Search our genomes for your DNA or protein sequence". This section is highlighted with a red box.
- Variant Effect Predictor >**: A link to "Analyse your own variants and predict the functional consequences of known and unknown variants".

Below these sections is a search form with a dropdown menu set to "All species" and a text input field containing "e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease".

The main content area is divided into two main sections:

- All genomes**: A section where users can select a species. It currently shows "Pig breeds" (highlighted with a red box) and "View full list of all species".
- Favourite genomes**: A list of genomes with their latest versions:
  - Human: GRCh38.p14
  - Mouse: GRCm39
  - Zebrafish: GRCz11

To the right of these sections is a "Ensembl Rapid Release" box:

- 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.
- The box contains a "Go" button and a link "Rapid Release news" on our blog.

At the bottom of the page are several links to other Ensembl features:

- Compare genes across species
- Find SNPs and other variants for my gene
- Gene expression in different tissues
- Retrieve gene sequence
- Find a Data Display
- Use my own data in Ensembl

At the very bottom, there is footer information about Ensembl's mission, funding, and acknowledgments, along with logos for EMBL-EBI, Global Coordinated Data Resource, and elair Core Data Resource.

Ensembl数据库中，Human的基因组的基本信息。

包括了近2万的编码基因，2万+的非编码基因以及1w+的假基因。



## Human assembly and gene annotation

### Assembly

#### Statistics

#### Summary

Assembly	GRCh38.p14 (Genome Reference Consortium Human Build 38), INSDC Assembly <a href="#">GCA_000001405.29</a> , Dec 2013
Base Pairs	3,099,750,718
Golden Path Length	3,099,750,718

#### Gene counts (Primary assembly)

Coding genes	19,831 (excl 650 readthrough)
Non coding genes	25,959
Small non coding genes	4,864
Long non coding genes	18,874 (excl 319 readthrough)
Misc non coding genes	2,221
Pseudogenes	15,239 (excl 1 readthrough)
Gene transcripts	252,894

在科研人员完成项目时，也会有科研人员将项目相关的数据上传到EBI的数据库，在ENA数据库中也能够直接搜索。通过不同数据库上传的Study Accession号会有区别。

The screenshot shows the ENA homepage with a search bar containing 'lung'. Below the search bar, there are two sections: 'Search results for lung' and a table of study records.

**Search results for lung**

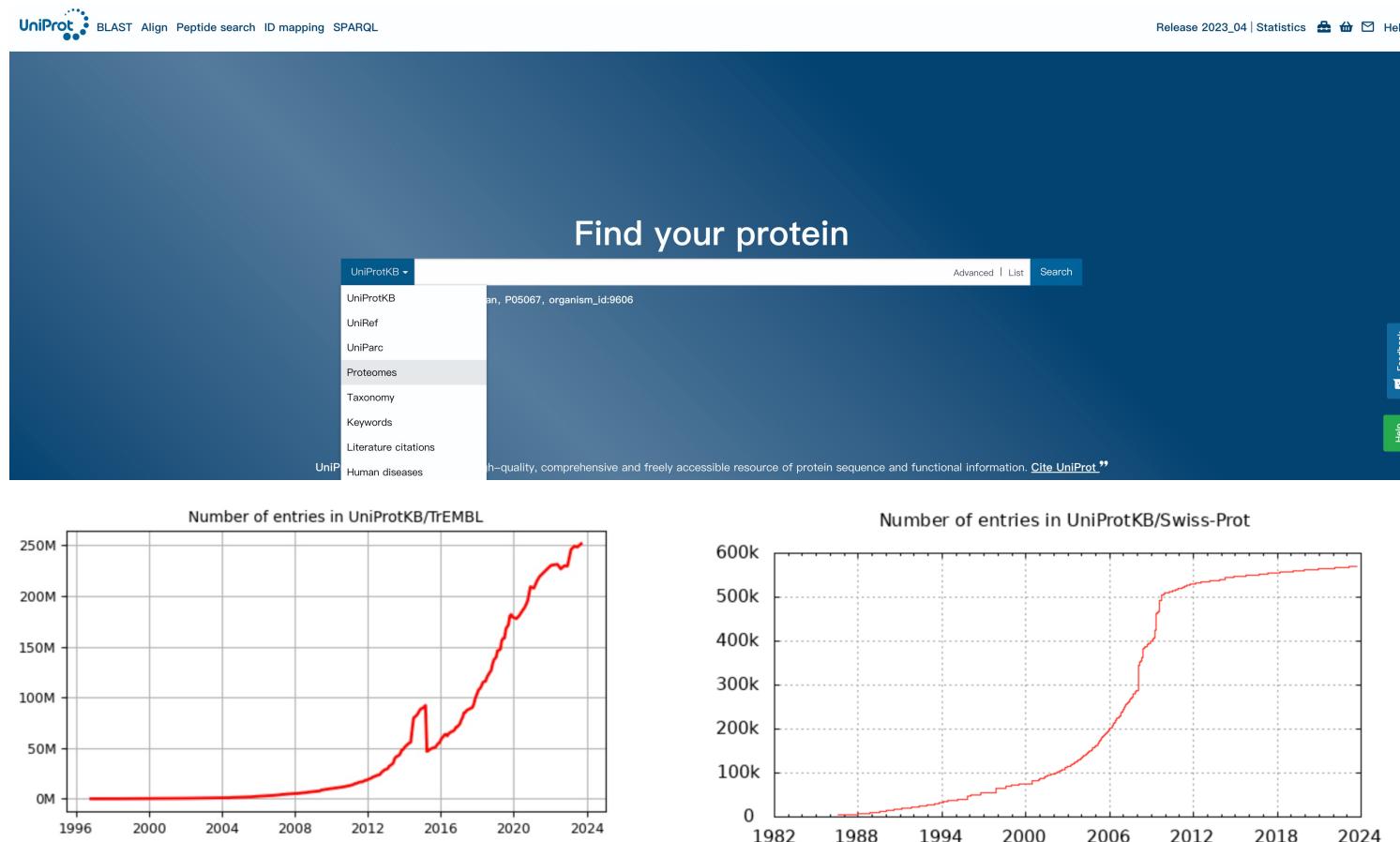
- Assembly
  - Assembly (1,843)
- Sequence
  - Sequence (186,403)
  - Sequence (Standard) (186,403)
- Contig set
  - Genome assembly contig set (2,122)
  - Transcriptome assembly contig set (2)
  - Targeted locus study contig set (1)
- Coding
  - Coding (60,652)
  - Coding (CON) (14,191)
  - Coding (Standard) (4,329)

**Study**

Accession	Description/Title	Download ENA records:
SRP006480	Viruses from cystic fibrosis lung tissue Metagenome	XML TSV
SRP114665	Effects of silver nanoparticles on mouse lung microbiome	
SRP331571	lung metagenome Raw sequence reads	
SRP336614	lung metagenome Raw sequence reads	
ERP106873	The human lung sputum eukaryotome as potential indicators of lung cancer	

UniProt数据库中主要存储的是蛋白质信息，可以搜索蛋白质组，疾病相关的蛋白等。

蛋白质主要包括两种，TrEMBL是未人工校对的，Swiss-Prot是人工校对过的，可以看到未经过校对的数据是近乎指数增长的。



# EMBL-EBI Quick GO

**GO (Gene Ontology)** 是一个标准化系统，用于描述跨物种的基因和基因产物属性，促进生物功能、过程和细胞成分的一致注释。QuickGO 简化并加速了访问和使用 GO 术语的过程。

The screenshot shows the Quick GO homepage with a blue header and a white main content area.

**Header:**

- Quick GO logo
- Gene Ontology and GO Annotations
- Navigation links: Help, Contact, API, Basket

**Main Content Area:**

- Search:** A large search bar with placeholder text "e.g apoptosis; GO:0006915; ECO:0000314; tropomyosin".
- View GO Annotations:** A section illustrating the relationship between a central entity ("Protein/Complex/RNA") and its annotations. Arrows point from the central circle to four categories: "enables" (molecular function), "involved in" (biological process), "part of" (cellular component), and "enables" (molecular function).
- Explore biology:** A section describing how to use GO terms (slims) to describe areas of interest. It shows a flow: "Choose your terms" leads to "Protein/Complex/RNA", which then leads to "Get GO slim annotations".

GO version 2023-10-12  
Annotation set created on 2023-09-19 07:23

# 网页在线基因注释工具 - Ami GO & Metascape

将感兴趣的基因list输入，就可以得到基因注释的结果。

AmiGO

The screenshot shows the AmiGO 2 homepage. At the top, there's a navigation bar with links for Home, Search, Browse, Tools & Resources, Help, Feedback, and About. Below the navigation is a search bar with a "Search" button. The main content area is divided into several sections:

- PubMed Search:** A form to search by PMID with a "Search" button.
- Search Templates:** A section for predefined templates with a "Go" button.
- Advanced Search:** An interactive search interface for Gene Ontology data with a "Search" button.
- Browse the Ontology:** A section for exploring the ontology structure with a "Go" button.
- Term Enrichment Service:** A tool for analyzing gene lists with a "Go" button.
- Statistics:** A section showing recent statistics about Gene Ontology data with a "Go" button.
- And Much More...**: A section listing other available tools with a "Go" button.

Metascape

The screenshot shows the Metascape web interface. At the top, there's a logo with the text "Metascape A Gene Annotation & Analysis Resource". The main area is divided into three steps:

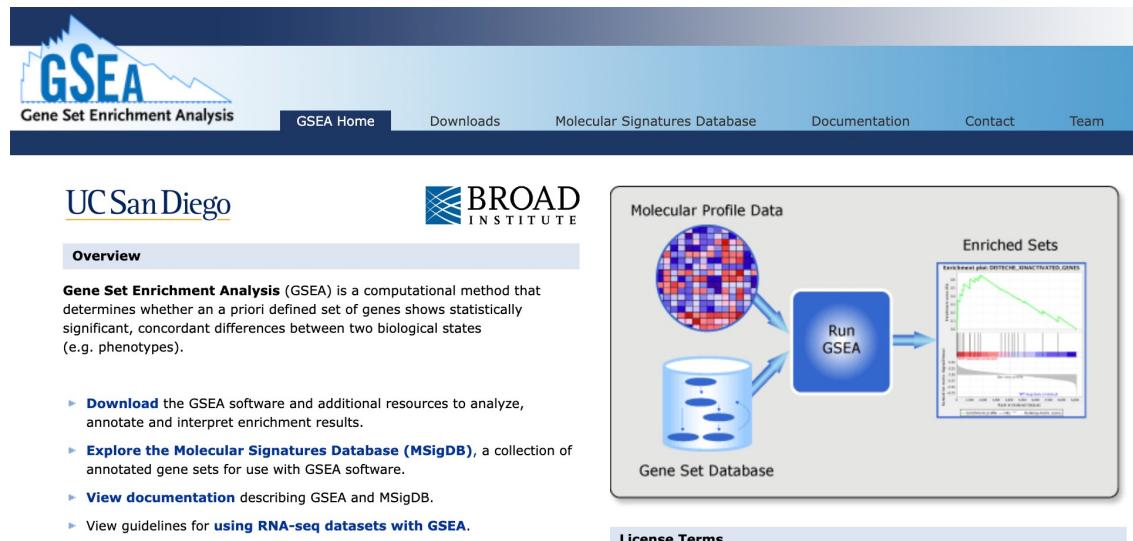
- Step 1: Multiple Gene Lists**  
A form to "Drag & drop your file (.xls, .xlsx, .csv)" or "Or paste a gene list" (Accept Gene ID/Symbol/RefSeq/Ensembl/UniProt/UCSC).  
Upload File Format:
  - Single List: .xls/.xlsx, .csv, .txt
  - Multiple List: .xls/.xlsx, .csv, .txt
- Step 2:** (partially visible)
- Step 3:** (partially visible)  
Buttons for "Express Analysis", "Custom Analysis", and "Batch Analysis?"

On the right side, there are two panels:

- News & Updates:** A purple panel listing recent releases and updates, such as "Current version v3.5.20230501" and "2021-12-18 Release MSBio".
- Message Board:** A green panel showing recent messages, such as "2023-09-04 Database updated to release 2023-09-01. MSBio updated accordingly." and "2023-06-22 ChatGPT-based gene annotation added (see blog)".

# 其他GO注释工具 – GSEA & MSigDB cluster profiler

## GSEA



## clusterProfiler

[PDF] [clusterProfiler 4.0: A universal enrichment tool for interpreting omics data](#)

T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng... - The innovation, 2021 - cell.com

... version of our popular Bioconductor package, [clusterProfiler](#) 4.0. This package has been ...

We anticipate that [clusterProfiler](#) 4.0 will be applied to a wide range of scenarios across ...

☆ 保存 引用 被引用次数: 3283 相关文章 所有 7 个版本

[clusterProfiler: an R package for comparing biological themes among gene clusters](#)

G Yu, LG Wang, Y Han, QY He - Omics: a journal of integrative ..., 2012 - liebertpub.com

... Here, we present an R package, [clusterProfiler](#) that automates the process of biological-...

Here, we present an R package called [clusterProfiler](#) for statistical analysis of GO and KEGG...

☆ 保存 引用 被引用次数: 18955 相关文章 所有 9 个版本

# 相关数据库

在科研工作中，与功能基因组相关的数据也会上传到ENCODE数据库。

Sections	Figures	References
<a href="#">Abstract</a>		
<a href="#">Main</a>		
<a href="#">Technology workflow and data quality</a>		
<a href="#">Spatial comapping of mouse embryo</a>		
<a href="#">Spatial ATAC–RNA-seq of mouse brain</a>		
<a href="#">Spatial CUT&amp;Tag–RNA-seq of mouse brain</a>		
<a href="#">Region-specific gene expression regulation</a>		
<a href="#">Spatial comapping of human brain</a>		
<a href="#">Discussion</a>		
<a href="#">Methods</a>		
<a href="#">Data availability</a>		
<a href="#">Code availability</a>		

## Data availability

Raw and processed data reported in this paper are deposited in the Gene Expression Omnibus with accession code [GSE205055](#). These datasets are available as web resources and can be browsed within the tissue spatial coordinates in the UCSC Cell and Genome Browser (<https://brain-spatial-omics.cells.ucsc.edu>), and in our own data portal generated with AtlasXplore (<https://web.atlasxomics.com/visualization/Fan>). Data are also available at <https://ki.se/en/mgb/oligointernode>. The resulting fastq files were aligned to either the human reference genome (GRCh38) (<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/>) or mouse reference genome (GRCm38) (<https://hgdownload.soe.ucsc.edu/goldenPath/mm10/chromosomes/>).

Published data for integration and quality comparison are available online: ENCODE ATAC-seq (E13.5 mouse embryo) (<https://www.encodeproject.org/search/?>)

Zhang D, Deng Y, Kukanja P, et al. *Nature*, 2023.

ENCODE数据库是NIH下属的National Human Genome Research Institute (NHGRI)发起的，主要是表征功能基因组学。

The screenshot shows the ENCODE Experiment search interface. On the left, there is a sidebar with various filters: Assay type (selected 'Assay'), Biosample, Library, Analysis, Provenance, Quality, and Other filters. The 'Assay type' dropdown is highlighted with a red box. The main area displays 8 results for ATAC-seq experiments:

- ATAC-seq in liver**: Mus musculus strain B6NCrl liver tissue embryo (13.5 days).  
Lab: Bing Ren, UCSD  
Project: ENCODE  
Reference Epigenome: ENCSR266TXT  
Organism Development Series: ENCSR326DKM
- ATAC-seq in heart**: Mus musculus strain B6NCrl heart tissue embryo (13.5 days).  
Lab: Bing Ren, UCSD  
Project: ENCODE  
Reference Epigenome: ENCSR107EDN  
Organism Development Series: ENCSR028ALM
- ATAC-seq in embryonic facial prominence**: Mus musculus strain B6NCrl embryonic facial prominence tissue embryo (13.5 days).  
Lab: Bing Ren, UCSD  
Project: ENCODE  
Reference Epigenome: ENCSR792SPH  
Organism Development Series: ENCSR660CNU
- ATAC-seq in limb**: Mus musculus strain B6NCrl limb tissue embryo (13.5 days).  
Lab: Bing Ren, UCSD  
Project: ENCODE  
Reference Epigenome: ENCSR424HQH  
Organism Development Series: ENCSR780UFT
- ATAC-seq in neural tube**: Mus musculus strain B6NCrl neural tube tissue embryo (13.5 days).  
Lab: Bing Ren, UCSD  
Project: ENCODE

Each result card includes an 'Experiment' button, a 'released' status indicator, and a small icon showing the count of data types (red triangle for RNA, orange square for ChIP, yellow circle for ATAC).

可以通过ENCODE数据库，查看实验具体技术，样本物种、建库方法等，同时提供了基因组可视化功能。

## Experiment summary for ENCSR343TXK

doi:10.17989/ENCSR343TXK

4

Summary		Attribution	
Status:	released	Lab:	Bing Ren, UCSD
Assay:	ATAC-seq	Award:	U54HG006997 (Bing Ren, UCSD)
Biosample summary:	<i>Mus musculus</i> strain B6NCrl liver tissue embryo (13.5 days)	Project:	ENCODE
Biosample Type:	tissue	External resources:	<a href="#">GEO:GSE172763</a>
Replication type:	isogenic	Aliases:	bing-ren:e13.5_liver_ATAC-seq
Description:	ATAC-seq on embryonic e13.5 mouse liver	Date submitted:	March 14, 2017
Nucleic acid type:	DNA	Date released:	October 10, 2017
Size range:	200-750	Reference:	<a href="#">ENCSR266TXT</a>
Lysis method:	ATAC buffer	Epigenome:	
Fragmentation methods:	chemical (Nextera tagmentation)	Organism Development Series:	<a href="#">ENCSR326DKM</a>
Size selection method:	SPRI beads	Annotation (gkmSVM-model):	<a href="#">ENCSR929QY</a>
Platform:	Illumina HiSeq 4000	Tags:	<a href="#">ENCODE ENCYCLOPEDIA</a>

### Files

Choose analysis: ENCODE4 v1.8.0 mm10

Filter files

File format: bigBed narrowPeak 8, bigWig 6

Output type: signal p-value 3, IDR thresholded peaks 3, pseudoreplicated peaks 3, fold change over control 3, conservative IDR thresholded peaks 1, replicated peaks 1

Replicates: 1, 2 6, 1 4, 2 4

Genome browser, Association graph, File details, Include deprecated files

Search for a gene: Enter gene name here

Sort by: Replicates, Output type, Reset coordinates

chr12 56680700bp to 56803100bp

mm10, GENE M21, Pax9, Gm13524, Slc25a21, representative DNase hypersensitivity sites, cCRE, all, ENCF0650RDT (rep 1, 2), ENCF0715YB (rep 1, 2), ENCF0421VAC (rep 1, 2), ENCF0583YQU (rep 1, 2)

Legend: Loading, Loading, Loading, Loading, Loading, Loading

# UCSC & WashU Genome Browser

常用的数据库还包括提供可视化功能的网页，如2000年推出的UCSC Genome Browser，同样的，在UCSC genome browser。也可以下载常见的基因组数据。

The screenshot shows the UCSC Genome Browser homepage. At the top, there's a navigation bar with links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Projects, Help, and About Us. Below the navigation bar is a large, colorful visualization showing genomic tracks for genes like ATP1B2, TP53, and WRAP53. To the left of the visualization, a sidebar lists various genome assemblies: Human GRCh38/hg38, Human GRCh37/hg19, Human T2T-CHM13/hs1, Mouse GRCm39/mm39, Mouse GRCm38/mm10, Genome Archive GenArk, SARS-CoV-2 (COVID-19), and Other. A red box highlights the "Tools" link in the navigation bar. On the far left, there are icons for hg38, hg19, and mm39, each with a corresponding smiley face icon. Below the visualization, there's a "News" section with a list of recent updates. At the bottom right, there are "More news..." and "Subscribe" buttons.

- Genome Browser - Interactively visualize genomic data
- BLAT - Rapidly align sequences to the genome
- In-Silico PCR - Rapidly align PCR primer pairs to the genome
- Table Browser - Download and filter data from the Genome Browser
- LiftOver - Convert genome coordinates between assemblies
- REST API - Returns data requested in JSON format
- Variant Annotation Integrator - Annotate genomic variants
- More tools...

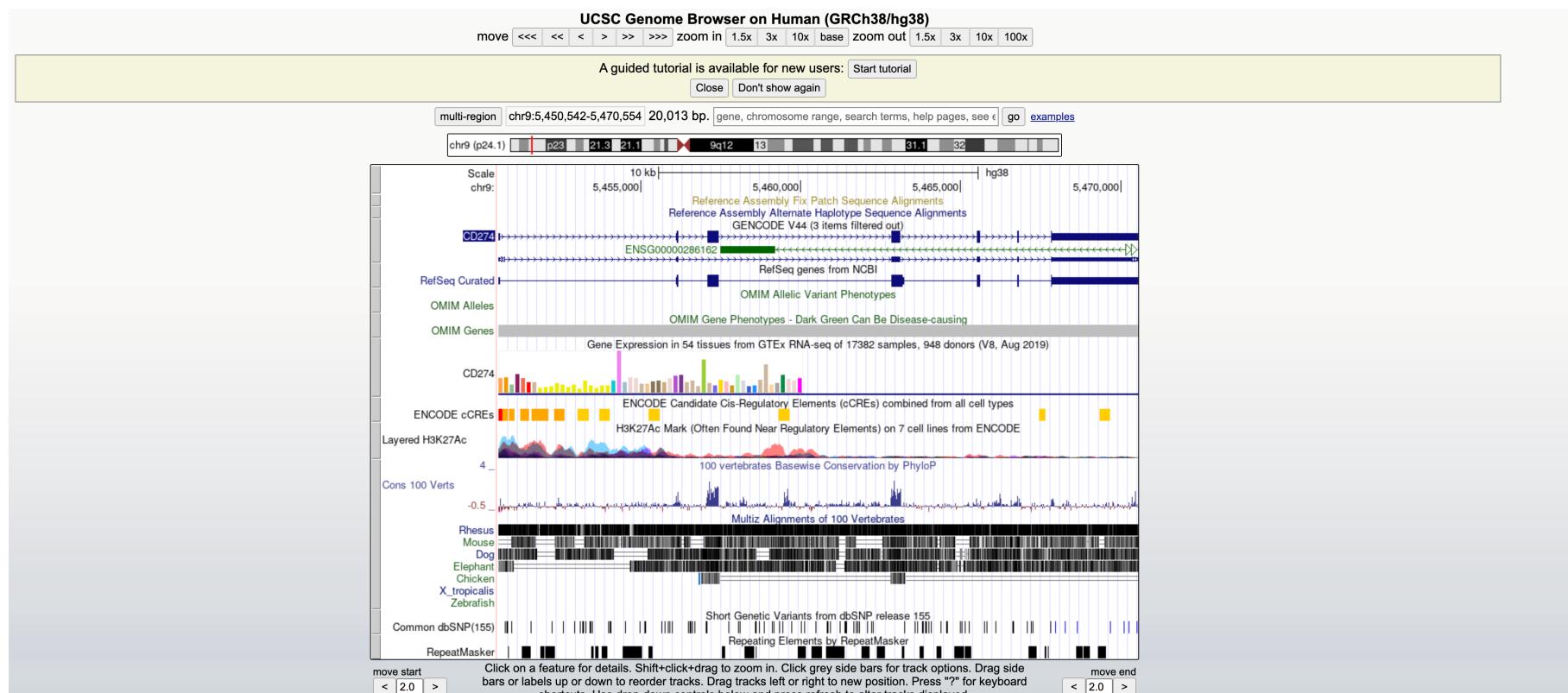
News

- Sep. 19, 2023 - EVA SNP release 5 for 36 assemblies
- Sep. 15, 2023 - New COSMIC Track for hg38
- Sep. 07, 2023 - New GENCODE "KnownGene" V44 (hg38) and VM33 (mm39)
- Aug. 18, 2023 - New GENCODE gene tracks: V44 (hg19/hg38) - VM33 (mm39)
- Aug. 07, 2023 - Introducing an interactive tutorial for the UCSC Genome Browser
- Aug. 01, 2023 - New ability to create duplicate tracks

More news...    Subscribe

# UCSC & WashU Genome Browser

在UCSC genome browser中，可以直接搜索基因或者染色质位置定位，能够看到数据库对应的数据的可视化结果。



# UCSC & WashU Genome Browser

其中包括了NCBI等数据库的数据，可以自己选择是否展示track。

**Mapping and Sequencing**

refresh

Base Position	P14 Fix Patches	P14 Alt Haplotypes	Assembly	Centromeres	Chromosome Band	Clone Ends	Exome Probesets
dense	pack	pack	hide	hide	hide	hide	hide
18 FISH Clones	Gap	GC Percent	GRC Contigs	GRC Incident	Hg19 Diff	INSDC	LiftOver & ReMap
hide	hide	hide	hide	hide	hide	hide	hide
LRG Regions	Mappability	New Problematic Regions	Recomb Rate	RefSeq Acc	Restr Enzymes	Scaffolds	Short Match
hide	hide	hide	hide	hide	hide	hide	hide
STS Markers							

**Genes and Gene Predictions**

refresh

Updated GENCODE V44	NCBI RefSeq	CCDS	CRISPR Targets	Updated GENCODE Versions	HGNC	IKMC Genes Mapped	LRG Transcripts
pack	dense	hide	hide	hide	hide	hide	hide
MANE	MGC Genes	Non-coding RNA	Old UCSC Genes	ORFeome Clones	Other RefSeq	Pfam in GENCODE	Prediction Archive
hide	hide	hide	hide	hide	hide	hide	hide
RetroGenes V9	TransMap V5	UCSC Alt Events	UniProt	hide	hide	hide	hide
hide	hide	hide	hide				

**Phenotype and Literature**

refresh

OMIM Alleles	CADD	Cancer Gene Expr	ClinGen	* ClinGen CNVs	ClinVar Variants	Constraint scores	Coriell CNVs
dense	hide	hide	hide	hide	hide	hide	hide
New COSMIC	* COSMIC Regions	DECIPHER CNVs	DECIPHER SNVs	Development Delay	GenCC	Gene Interactions	GeneReviews
hide	hide	hide	hide	hide	hide	hide	hide
GWAS Catalog	HGMD public	LOVD Variants	OMIM Cyto Loci	OMIM Genes	Orphanet	PanelApp	REVEL Scores
hide	hide	hide	hide	dense	hide	hide	hide
SNPedia	TCGA Pan-Cancer	UniProt Variants	Variants in Papers	hide	hide	hide	hide
hide	hide	hide	hide				

**Single Cell RNA-seq**

refresh

Blood (PBMC) Hao	Colon Wang	Cortex Velmeshev	Cross Tissue Nuclei	* Fetal Gene Atlas	Heart Cell Atlas	Ileum Wang	Kidney Stewart
hide	hide	hide	hide	hide	hide	hide	hide
Liver MacParland	Lung Travaglini	Merged Cells	Muscle De Micheli	Pancreas Baron	Placenta Vento-Tormo	Rectum Wang	Skin Sole-Boldo
hide	hide	hide	hide	hide	hide	hide	hide
Tabula Sapiens							

**mRNA and EST**

refresh

Human ESTs	Human mRNAs	Other ESTs	Other mRNAs	SIB Alt-Splicing	Spliced ESTs		
hide	hide	hide	hide	hide	hide		

**Expression**

refresh

GTEX Gene V8	GTEX RNA-Seq Coverage	Affy Archive	EFDnew Promoters	GNF Atlas 2	* GTEX Gene	GTEX Transcript	GWIPS-viz Riboseq
pack	hide	hide	hide	hide	hide	hide	hide
miRNA Tissue Atlas							

**Regulation**

refresh

ENCODE cCREs	ENCODE Regulation	CpG Islands	New FANTOM5	GeneHancer	GTEX cis-eQTLs	Hi-C and Micro-C	JASPAR Transcription Factors
dense	show	hide	hide	hide	hide	hide	hide
OpenAccess	RefSeq ENS Elements	RefSeq ChIP-seq					

# UCSC & WashU Genome Browser

也可以选择上传自己实验产生的数据进行可视化，与公共数据库进行比较。

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

**Add Custom Tracks**

clade Mammal genome Human assembly Dec. 2011

Display your own data as custom annotation tracks in the browser. Data must be in [bigBed](#), [bigWig](#), [bigGenePred](#), [bigInteract](#), [bigLolly](#), [bigMaf](#), [bigPsl](#), [bigWig](#), [BAM](#), [barChart](#), [VCF](#), [BED](#), [narrowPeak](#), [Personal Genome SNP](#), [PSL](#), or [WIG](#) formats.

- You can paste just the URL to the file, without a "track" line, for bigBed, bigWig, bigGenePred, BAM and VCF.
- To configure the display, set [track](#) and [browser](#) line attributes as described in the [User's Guide](#).

Examples are [here](#). If you do not have web-accessible data storage available, please see the [Hosting](#) section of the Track Hub Help documentation.

Please note a much more efficient way to load data is to use [Track Hubs](#), which are loaded from the [Track Hubs Portal](#) found in the menu under My Data.

Paste URLs or data: Or upload: 选择文件 未选择任何文件

Optional track documentation: Or upload: 选择文件 未选择任何文件

Click [here](#) for an HTML document template that may be used for Genome Browser track descriptions.

Custom Tracks

My Sessions

Track Hubs

Track Collection Builder

Public Sessions

# UCSC & WashU Genome Browser

在2010年，生物学排名比较靠前的WashU同样推出了网页版的可视化工具，**WashU Epigenome browser**。

41  Washington University in St. Louis 79.7  Shortlist

④ St. Louis, United States

## [WashU epigenome browser update 2022](#)

[D Li, D Purushotham, JK Harrison, S Hsu...](#) - Nucleic acids ..., 2022 - academic.oup.com

... opportunities for modern-day genome [browsers](#). The [WashU Epigenome Browser](#) was invented in 2011 to enable interactive explorations of genomic data in a web [browser](#) format (13). ...

☆ 保存 ⚡ 引用 被引用次数: 47 相关文章 所有 12 个版本

451-500  University of California, Santa Cruz n/a  Shortlist

④ Santa Cruz, United States

## [The UCSC genome browser database: 2022 update](#)

[BT Lee, GP Barber, A Benet-Pagès...](#) - Nucleic acids ..., 2022 - academic.oup.com

... SARS-CoV-2 [genomes](#) in a global phylogenetic tree enabling researchers to view the context of emerging mutations in our SARS-CoV-2 [Genome Browser](#). Other new software focuses ...

☆ 保存 ⚡ 引用 被引用次数: 149 相关文章 所有 8 个版本

# UCSC & WashU Genome Browser

[nature](#) > [nature methods](#) > [correspondence](#) > [article](#)

Published: 29 November 2011

## The Human Epigenome Browser at Washington University

[Xin Zhou](#), [Brett Maricque](#), [Mingchao Xie](#), [Daofeng Li](#), [Vasavi Sundaram](#), [Eric A Martin](#), [Brian C Koebbe](#),  
[Cydney Nielsen](#), [Martin Hirst](#), [Peggy Farnham](#), [Robert M Kuhn](#), [Jingchun Zhu](#), [Ivan Smirnov](#), [W James Kent](#), [David Haussler](#), [Pamela A F Madden](#), [Joseph F Costello](#)✉ & [Ting Wang](#)✉

[Nature Methods](#) 8, 989–990 (2011) | [Cite this article](#)

3069 Accesses | 241 Citations | 17 Altmetric | [Metrics](#)



Ting Wang

Professor of Genetics and Computer Science and Engineering, Washington University in St ...

在 genetics.wustl.edu 的电子邮件经过验证

genomics epigenomics computational biology DNA methylation transposable elements

被引用次数: 31916

# UCSC & WashU Genome Browser

与UCSC genome browser类似，可以选择不同物种的基因组查看track。

WashU Epigenome Browser

CHOOSE A GENOME

LOAD A SESSION

Please select a genome

Search for a genome...

 Human hg19 hg38 t2t-chm13-v1.1 t2t-chm13-v2.0	 Chimp panTro6 panTro5 panTro4	 Gorilla gorGor4 gorGor3
 Gibbon nomLeu3	 Baboon papAnu2	 Rhesus rheMac10 rheMac8 rheMac3 rheMac2
 Marmoset calJac4 calJac3	 Cow bosTau8	 Sheep oviAri4

# UCSC & WashU Genome Browser

选择基因组后，也能够上传用户数据进行可视化比较。

