Biological Databases

生物信息学
助教-刘柯
助教-方明昊

课程概览

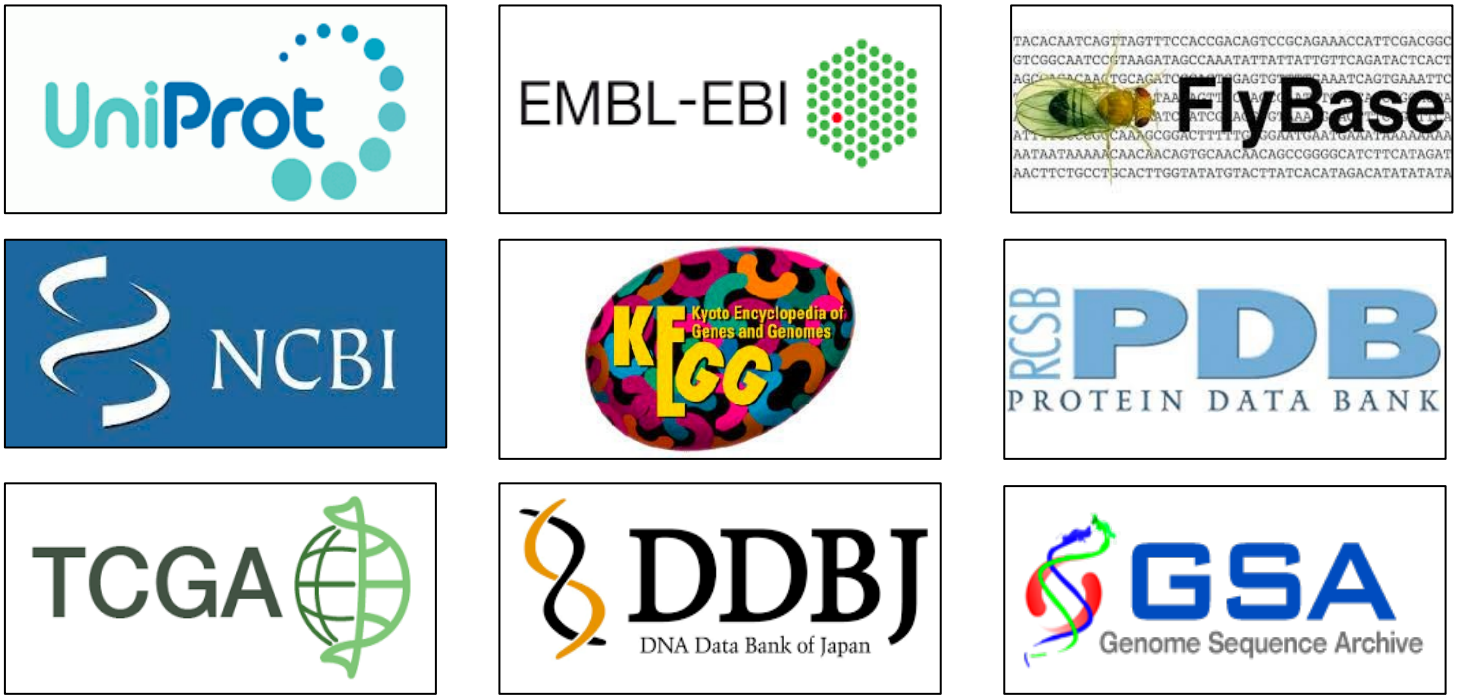
1 数据库基本概念

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

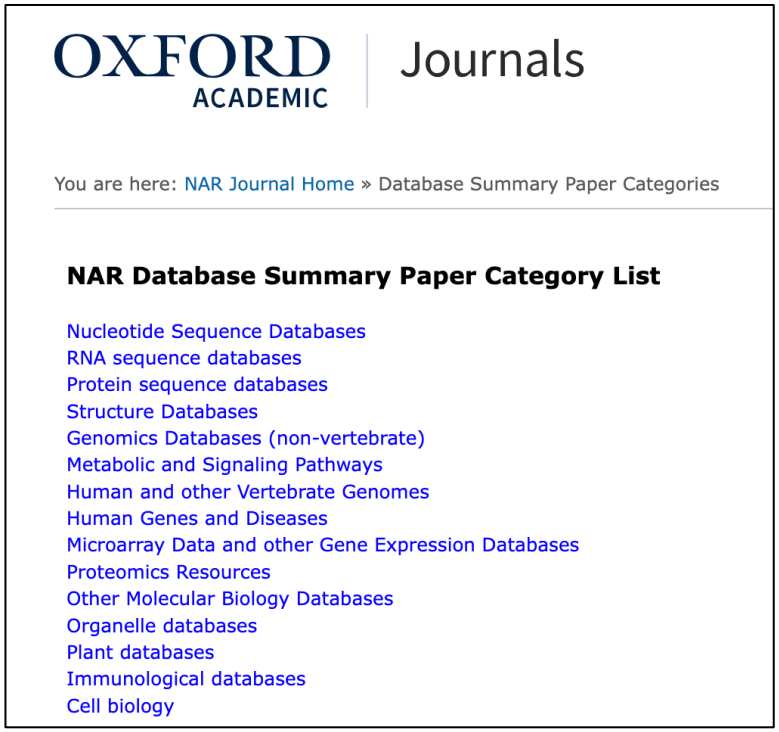
The 2023 *Nucleic Acids Research* Database Issue and the online molecular biology database collection

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


OXFORD ACADEMIC | Journals

You are here: NAR Journal Home » Database Summary Paper Categories

NAR Database Summary Paper Category List

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/

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

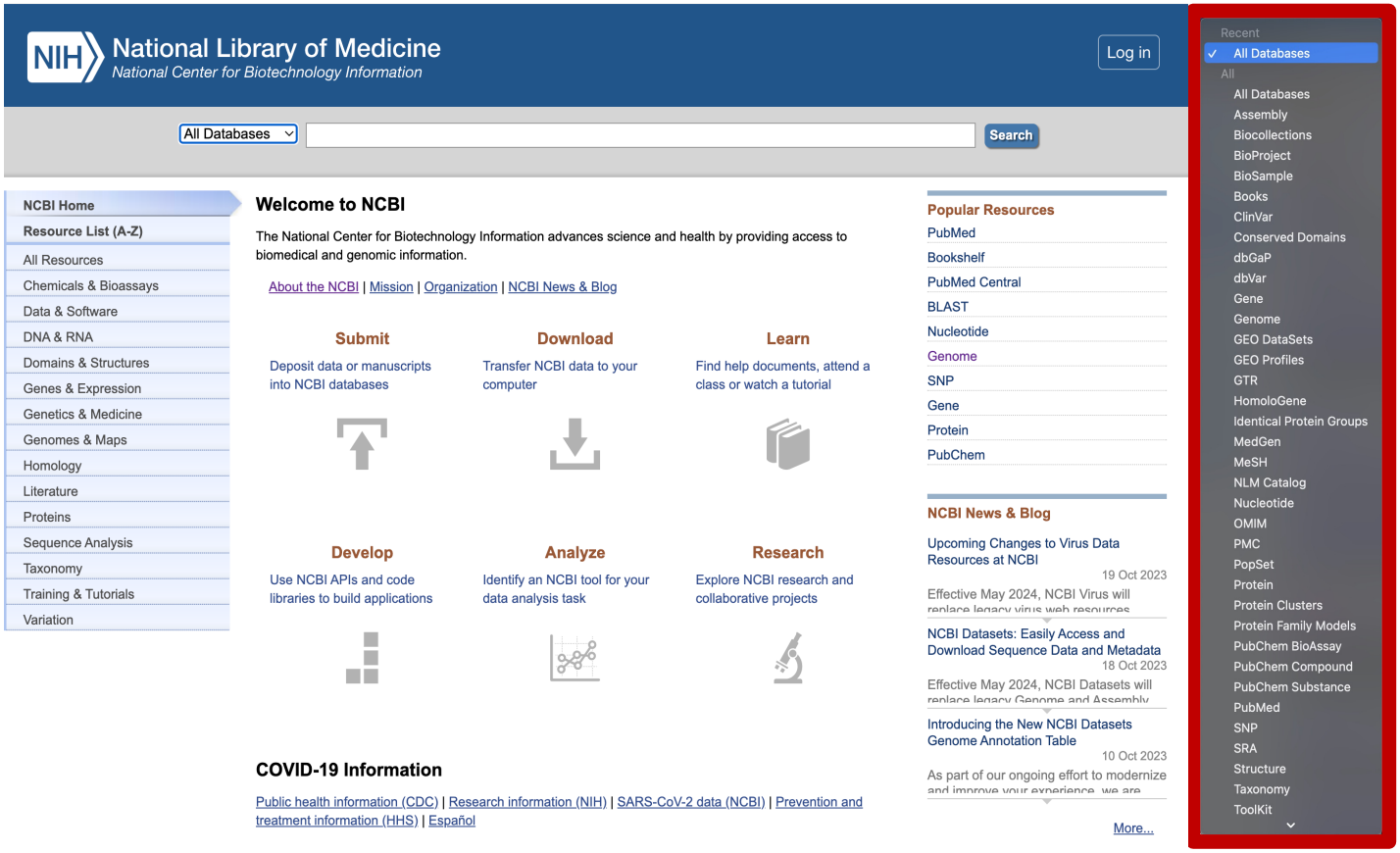
NCBI (National Center for Biotechnology Information)，隶属于NIH (National Institutes of Health)的NLB (National Library of Medicine)，受到的资助。NCBI成立于1988年，它的主要目标是满足对**生物和基因组数据的组织、存储和传播**日益增长的需求。

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.

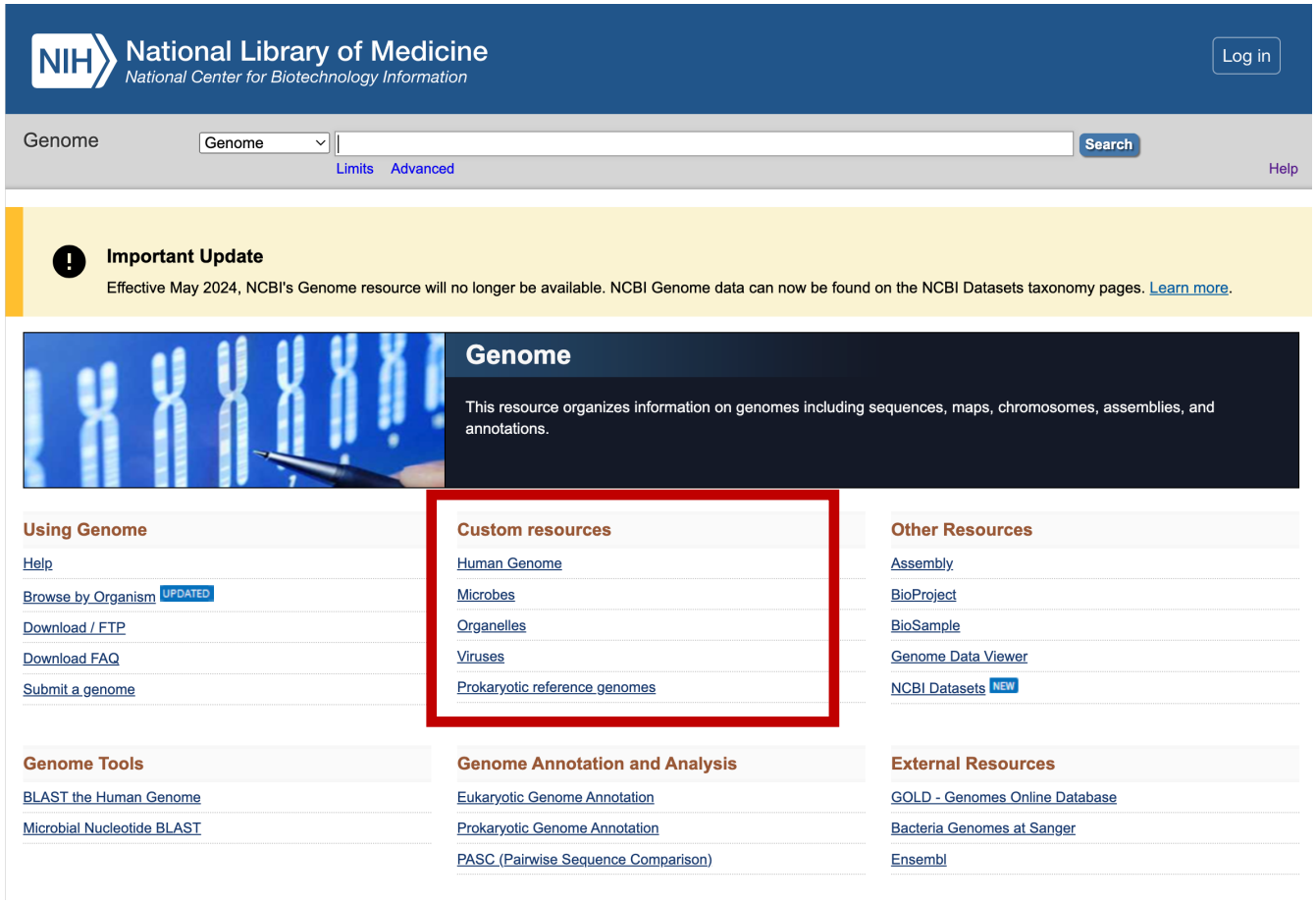
NCBI

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。

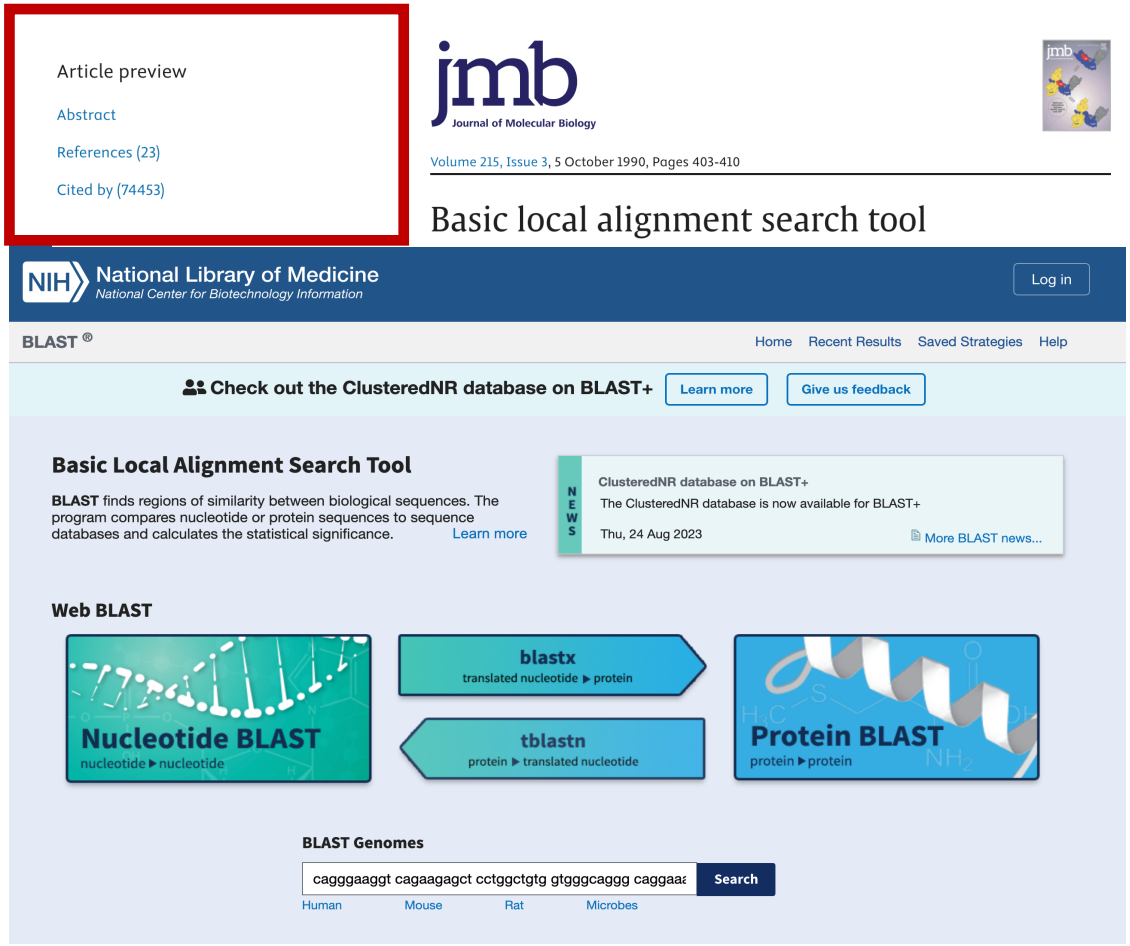


NCBI-genome

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。



NCBI-BLAST

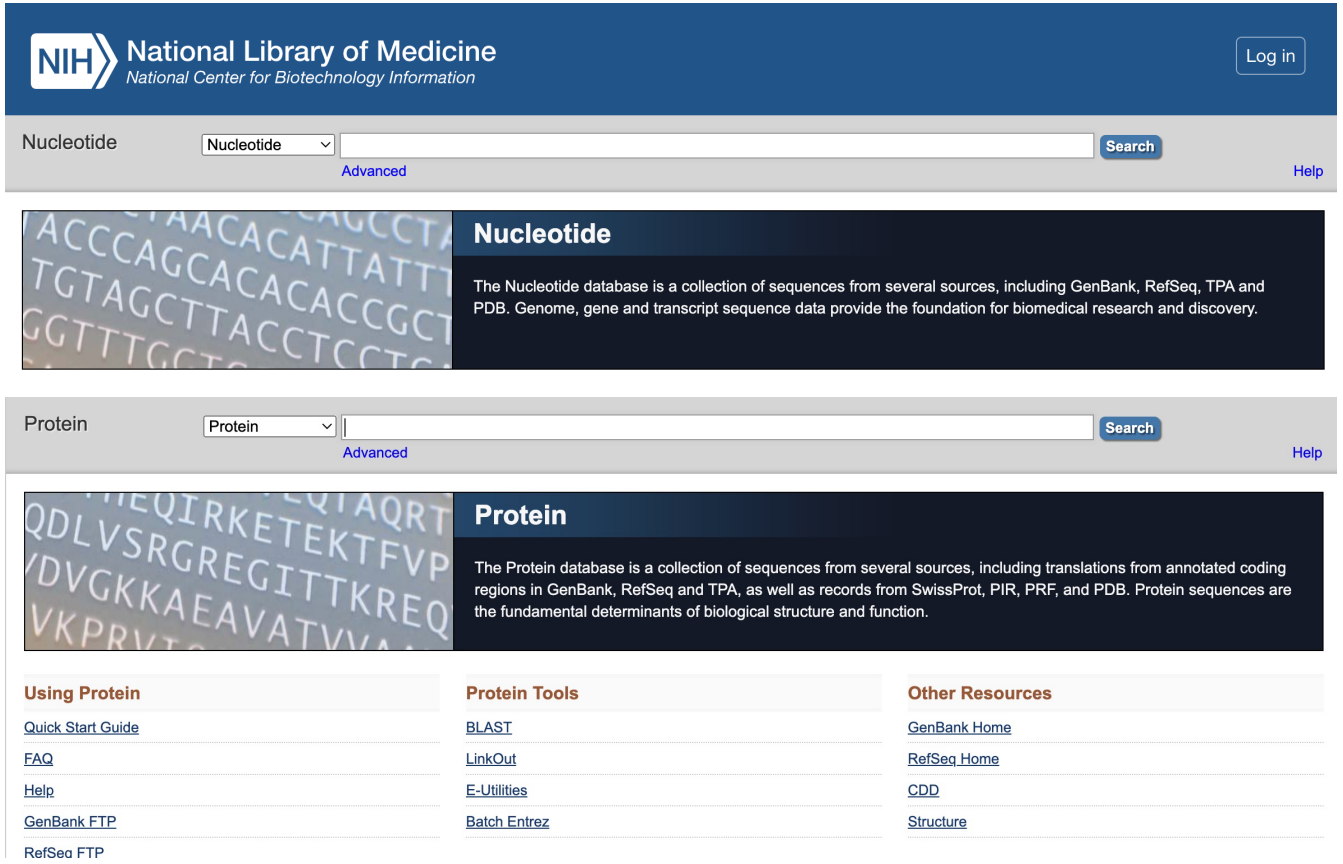


对于未知的序列，NCBI还提供了BLAST与NCBI的基因组进行比对。

BLAST从1990年发表至今，已被引用超过7万次。

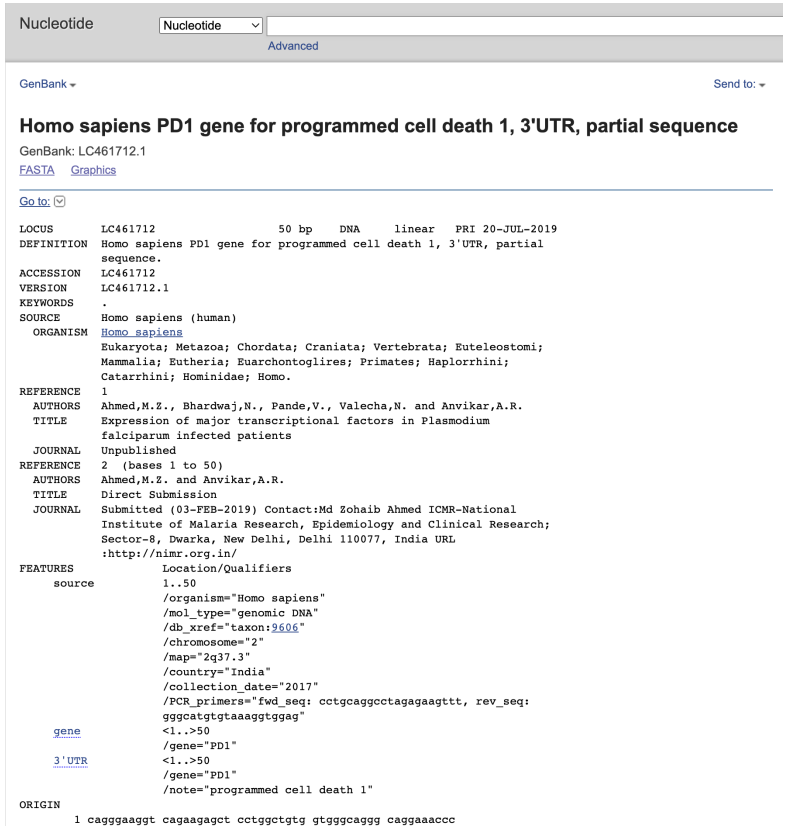
NCBI-RefSeq

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。



NCBI-RefSeq

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 | Send to:

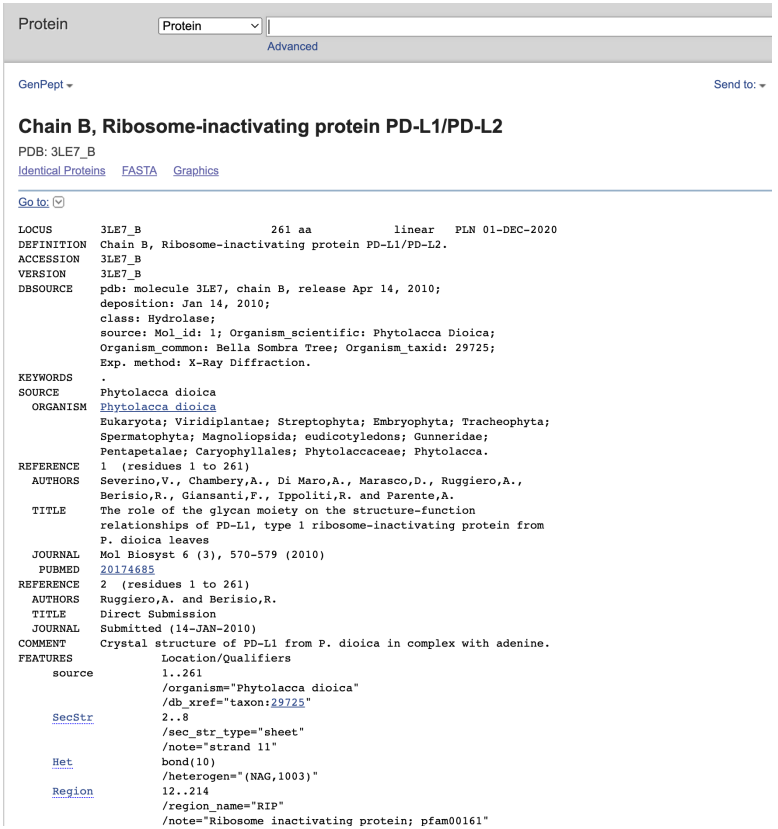
Homo sapiens PD1 gene for programmed cell death 1, 3'UTR, partial sequence

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: cctgcaggcctagagaagttt, rev_seq:
                     gggcatgtgtaaaggtggag"
     gene            <1..>50
                     /gene="PD1"
     3'UTR           <1..>50
                     /gene="PD1"
                     /note="programmed cell death 1"
ORIGIN
        1 cagggaaggt cagaagagct cctggctgtg gtgggcaggg caggaaaccc
```



Protein | Protein | Advanced

GenPept | Send to:

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

PDB: 3LE7_B

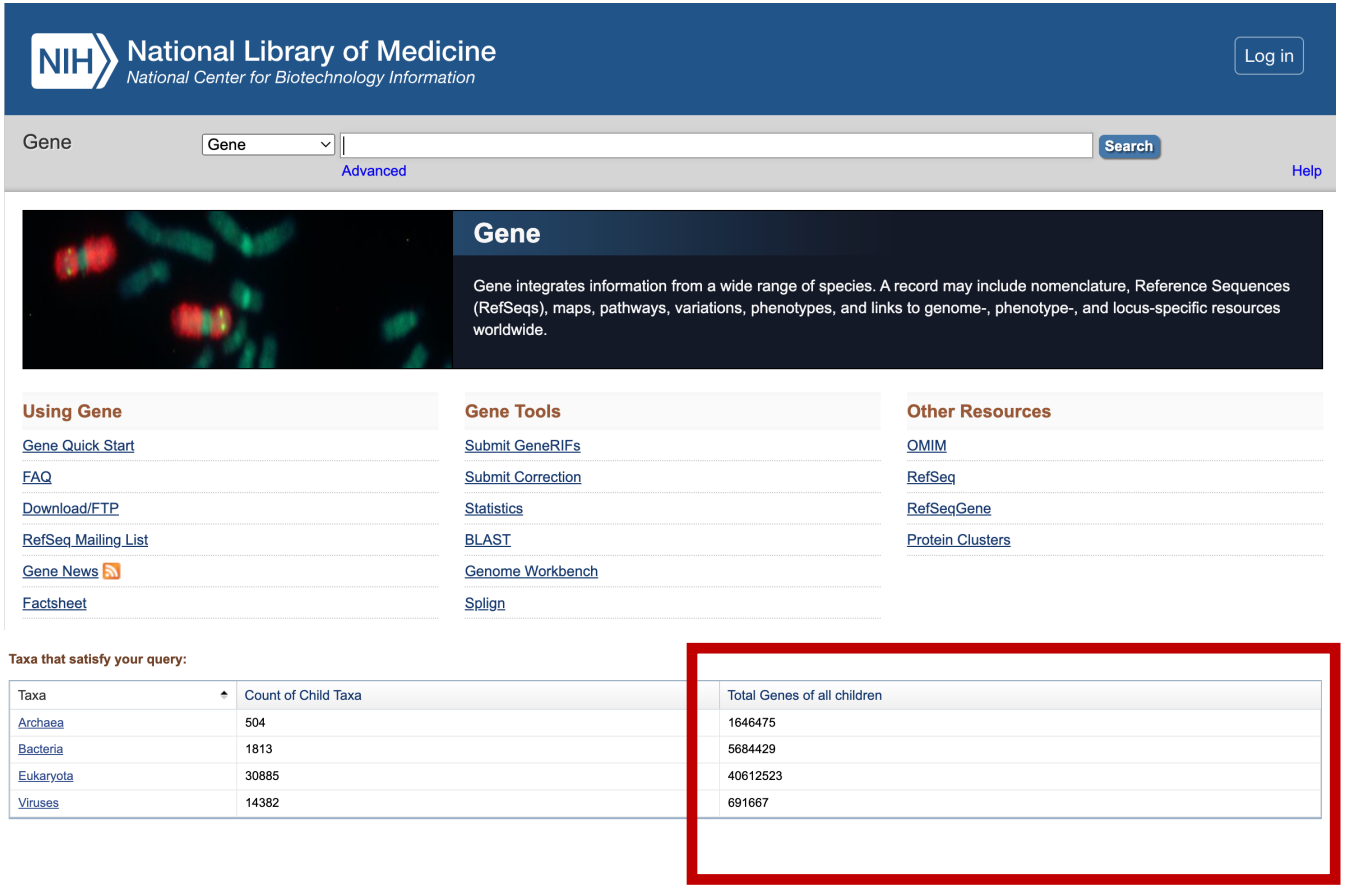
Identical Proteins FASTA Graphics

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 11"
     Het             bond(10)
                     /heterogen="(NAG,1003)"
     Region          12..214
                     /region_name="RIP"
                     /note="Ribosome inactivating protein; pfam00161"
```

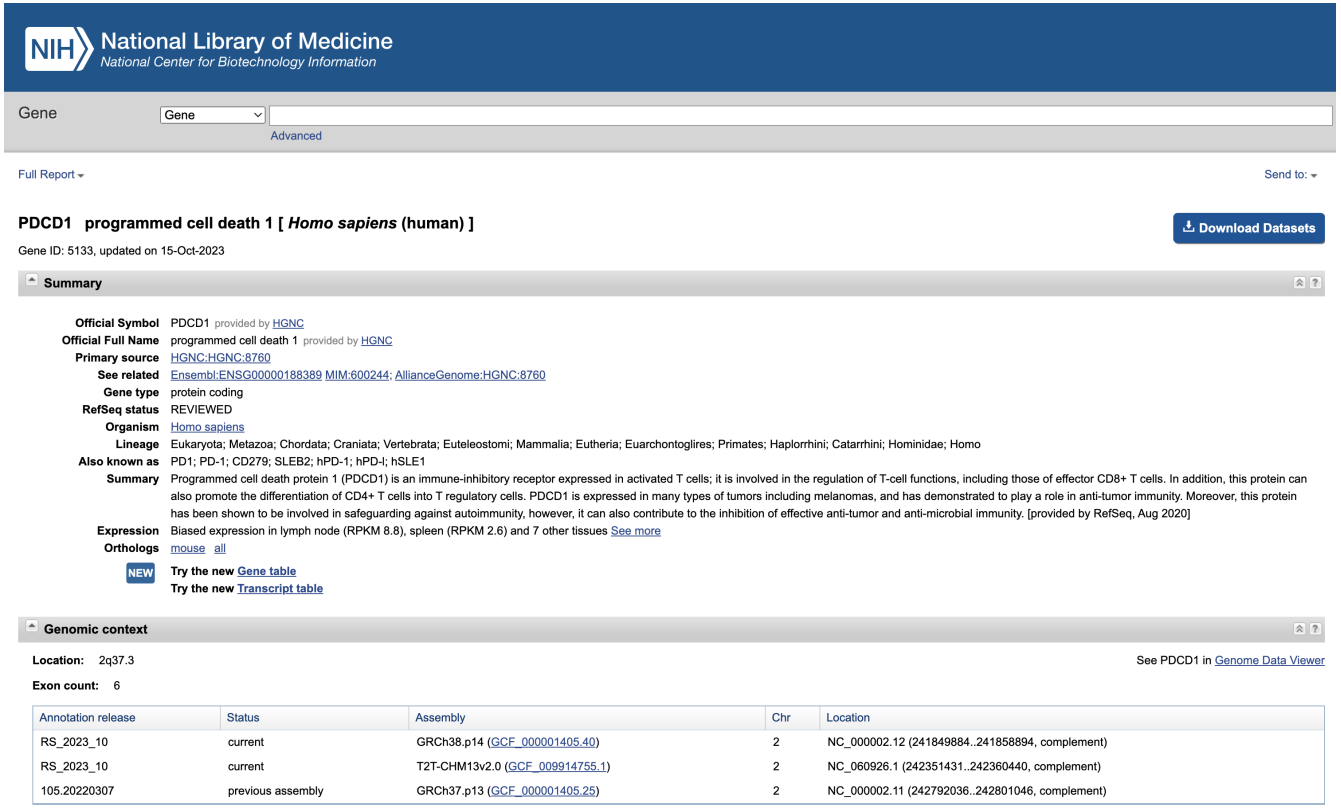
NCBI-Gene

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。



NCBI-Gene

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。



National Library of Medicine
National Center for Biotechnology Information

Gene | Gene | Advanced

Full Report | Send to:

PDCD1 programmed cell death 1 [*Homo sapiens* (human)]

Download Datasets

Gene ID: 5133, updated on 15-Oct-2023

Summary

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-l; 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

NEW **Try the new Gene table**
Try the new Transcript table

Genomic context

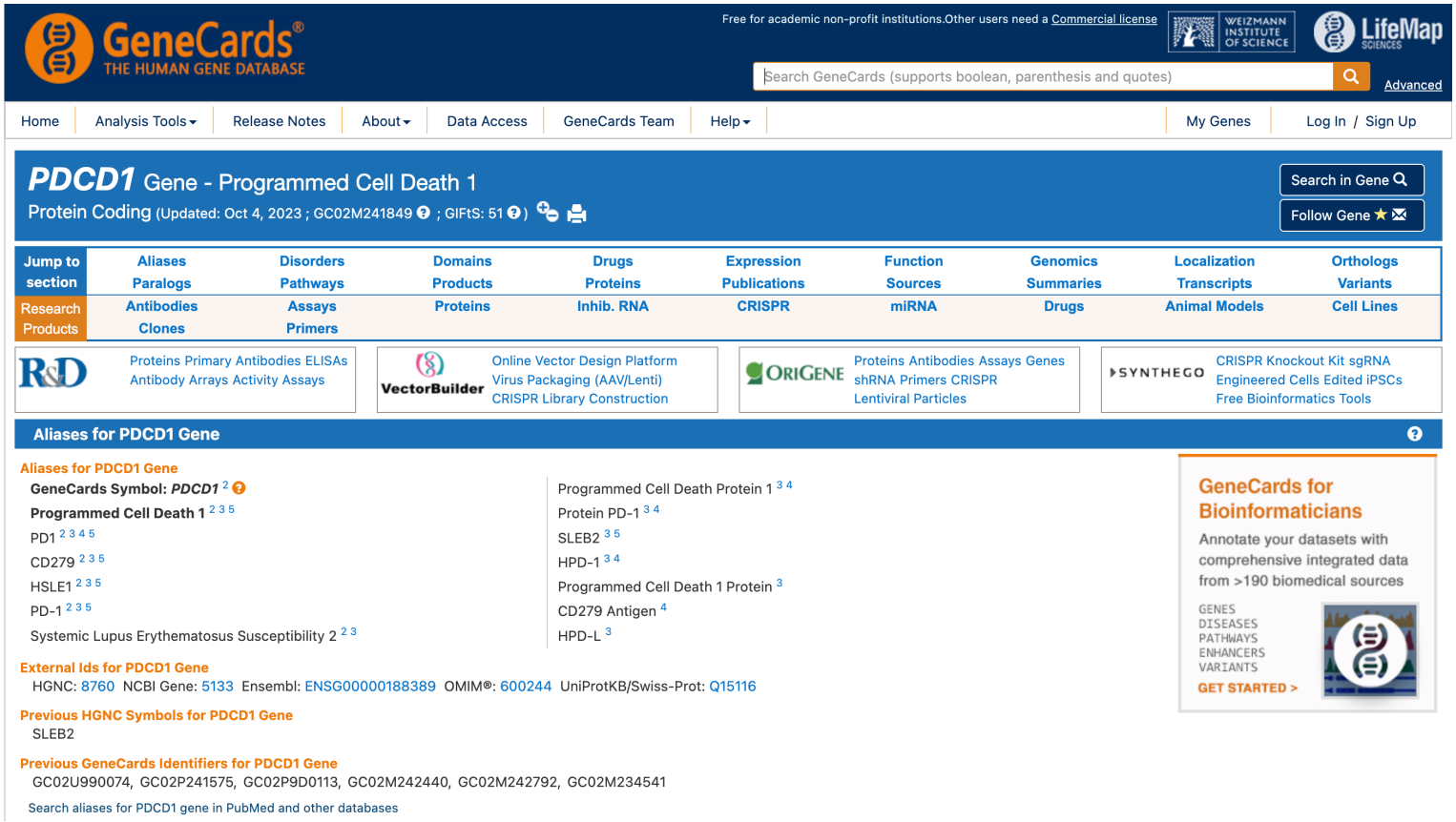
Location: 2q37.3 — See PDCD1 in Genome Data Viewer

Exon count: 6

Annotation release	Status	Assembly	Chr	Location
RS_2023_10	current	GRCh38.p14 (GCF_000001405.40)	2	NC_000002.12 (241849884..241858894, complement)
RS_2023_10	current	T2T-CHM13v2.0 (GCF_009914755.1)	2	NC_060926.1 (242351431..242360440, complement)
105.20220307	previous assembly	GRCh37.p13 (GCF_000001405.25)	2	NC_000002.11 (242792036..242801046, complement)

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。



NCBI-GEO

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。

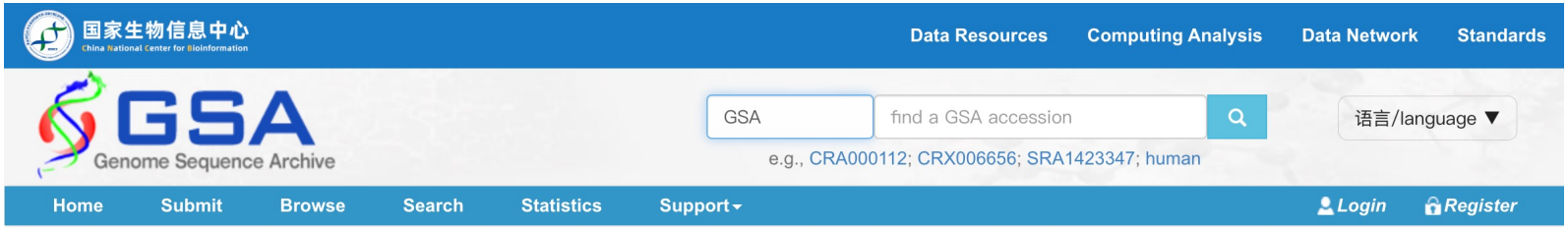
Sections | Figures | References

- 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.

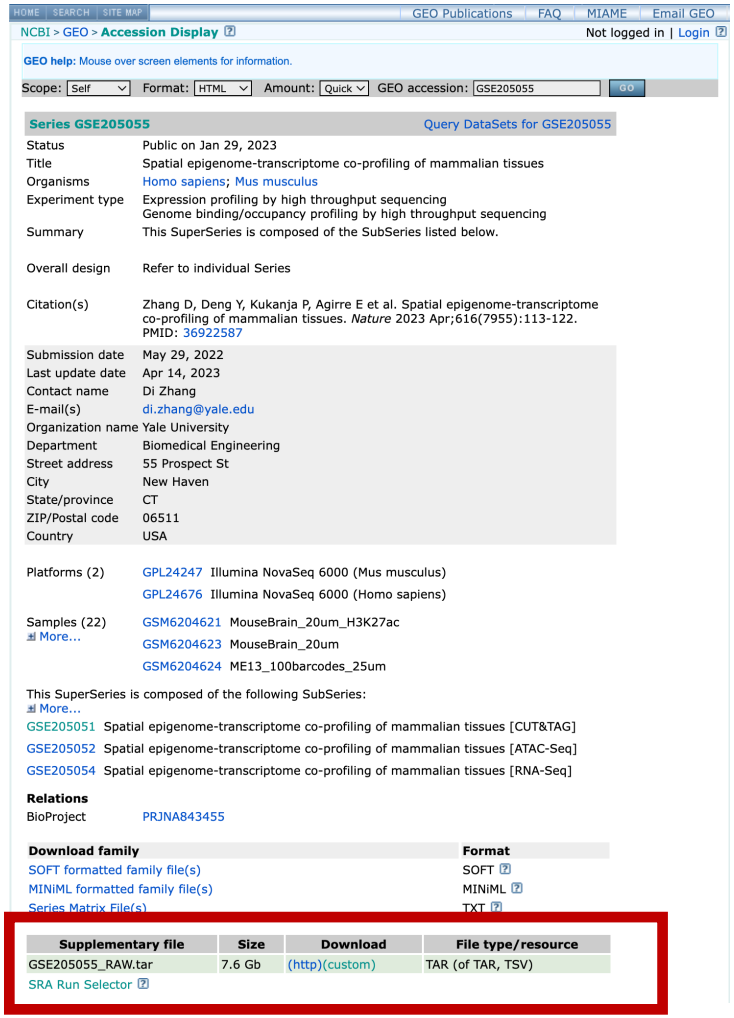


NCBI-GEO

在NCBI输入文章提供的GSE号，就能够搜索到对应的数据集。

NCBI-GEO

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 — GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display** — Not logged in | Login

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. *Nature* 2023 Apr;616(7955):113-122. PMID: 36922587
Submission date	May 29, 2022
Last update date	Apr 14, 2023
Contact name	Di Zhang
E-mail(s)	di.zhang@yale.edu
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)	GPL24247 Illumina NovaSeq 6000 (Mus musculus) GPL24676 Illumina NovaSeq 6000 (Homo sapiens)
Samples (22)	GSM6204621 MouseBrain_20um_H3K27ac GSM6204623 MouseBrain_20um GSM6204624 ME13_100barcodes_25um More...

This SuperSeries is composed of the following SubSeries:
More...
GSE205051 Spatial epigenome-transcriptome co-profiling of mammalian tissues [CUT&TAG]
GSE205052 Spatial epigenome-transcriptome co-profiling of mammalian tissues [ATAC-Seq]
GSE205054 Spatial epigenome-transcriptome co-profiling of mammalian tissues [RNA-Seq]

Relations
BioProject PRJNA843455

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE205055_RAW.tar	7.6 Gb	(http)(custom)	TAR (of TAR, TSV)

SRA Run Selector

Supplementary file	Size	Download	File type/resource
GSE205055_RAW.tar	7.6 Gb	(http)(custom)	TAR (of TAR, TSV)

SRA Run Selector

Custom GSE205055_RAW.tar archive:

Supplementary file	File size
GSM6204621_MouseBrain_20um_H3K27ac_fragments.tsv.gz	147.8 Mb
GSM6204621_MouseBrain_20um_H3K27ac_spatial.tar.gz	20.6 Mb
GSM6204623_MouseBrain_20um_fragments.tsv.gz	329.0 Mb
GSM6204623_MouseBrain_20um_spatial.tar.gz	17.7 Mb
GSM6204624_ME13_100barcodes_25um_fragments.tsv.gz	1.5 Gb
GSM6204624_ME13_100barcodes_25um_spatial.tar.gz	25.9 Mb
GSM6204635_MouseBrain_20um_H3K27ac_matrix.tsv.gz	4.0 Mb
GSM6204635_MouseBrain_20um_H3K27ac_spatial.tar.gz	20.6 Mb
GSM6204636_MouseBrain_20um_matrix.tsv.gz	3.6 Mb
GSM6204636_MouseBrain_20um_spatial.tar.gz	17.7 Mb
GSM6204637_ME13_100barcodes_25um_matrix.tsv.gz	7.9 Mb
GSM6204637_ME13_100barcodes_25um_spatial.tar.gz	25.9 Mb
GSM6206884_HumanBrain_50um_fragments.tsv.gz	435.3 Mb
GSM6206884_HumanBrain_50um_spatial.tar.gz	49.2 Mb
GSM6206885_HumanBrain_50um_matrix.tsv.gz	4.9 Mb
GSM6206885_HumanBrain_50um_spatial.tar.gz	49.2 Mb
GSM6704977_MouseBrain_20um_rep_H3K27ac_fragments.tsv.gz	158.2 Mb
GSM6704978_MouseBrain_20um_100barcodes_H3K27me3_fragments.tsv.gz	1.1 Gb
GSM6704979_MouseBrain_20um_100barcodes_H3K27ac_fragments.tsv.gz	1.0 Gb
GSM6704980_MouseBrain_20um_100barcodes_H3K4me3_fragments.tsv.gz	262.6 Mb

Select All | Cancel | Download | 0 file(s), 0 b

NCBI-SRA

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

Found 20 Items

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

NIH 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

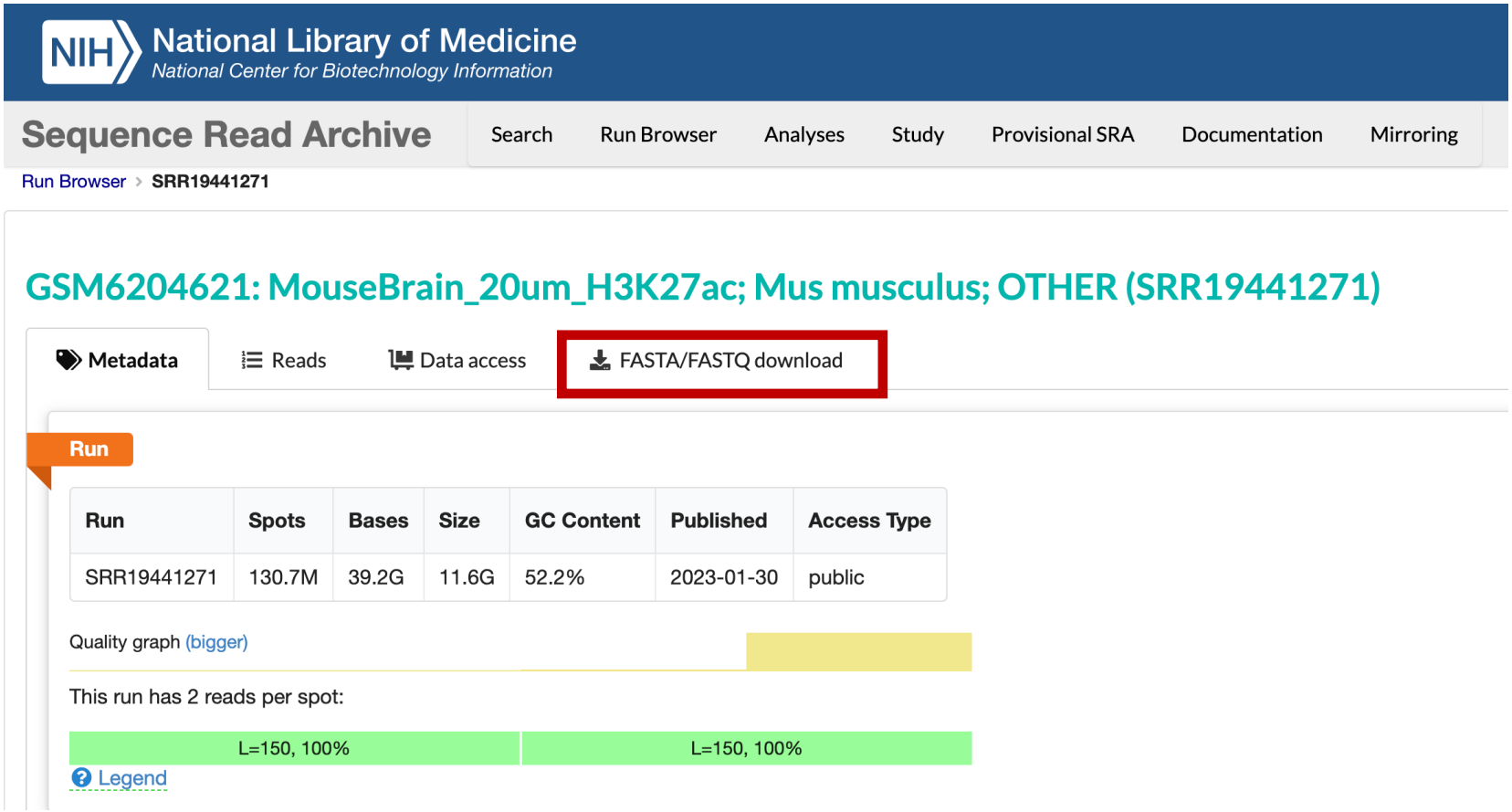
Library:
Name: GSM6204621
Instrument: Illumina NovaSeq 6000
Strategy: OTHER
Source: GENOMIC
Selection: other
Layout: PAIRED
Construction protocol: The fresh frozen tissue section was fixed with formaldehyde. The adapters loaded pA-Tn5 transposition containing a ligation linker was inserted into transposase 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 adapters 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 chromatin 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

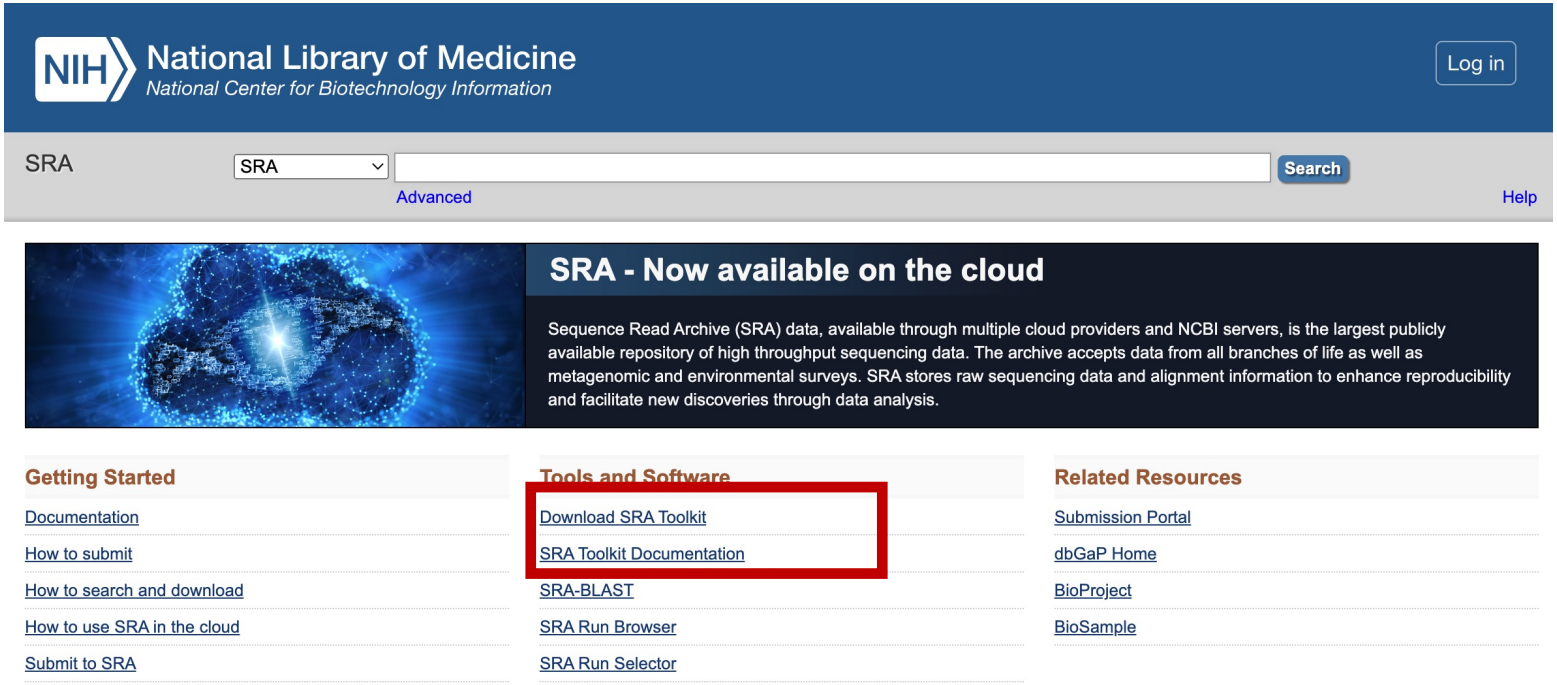
NCBI-SRA

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。



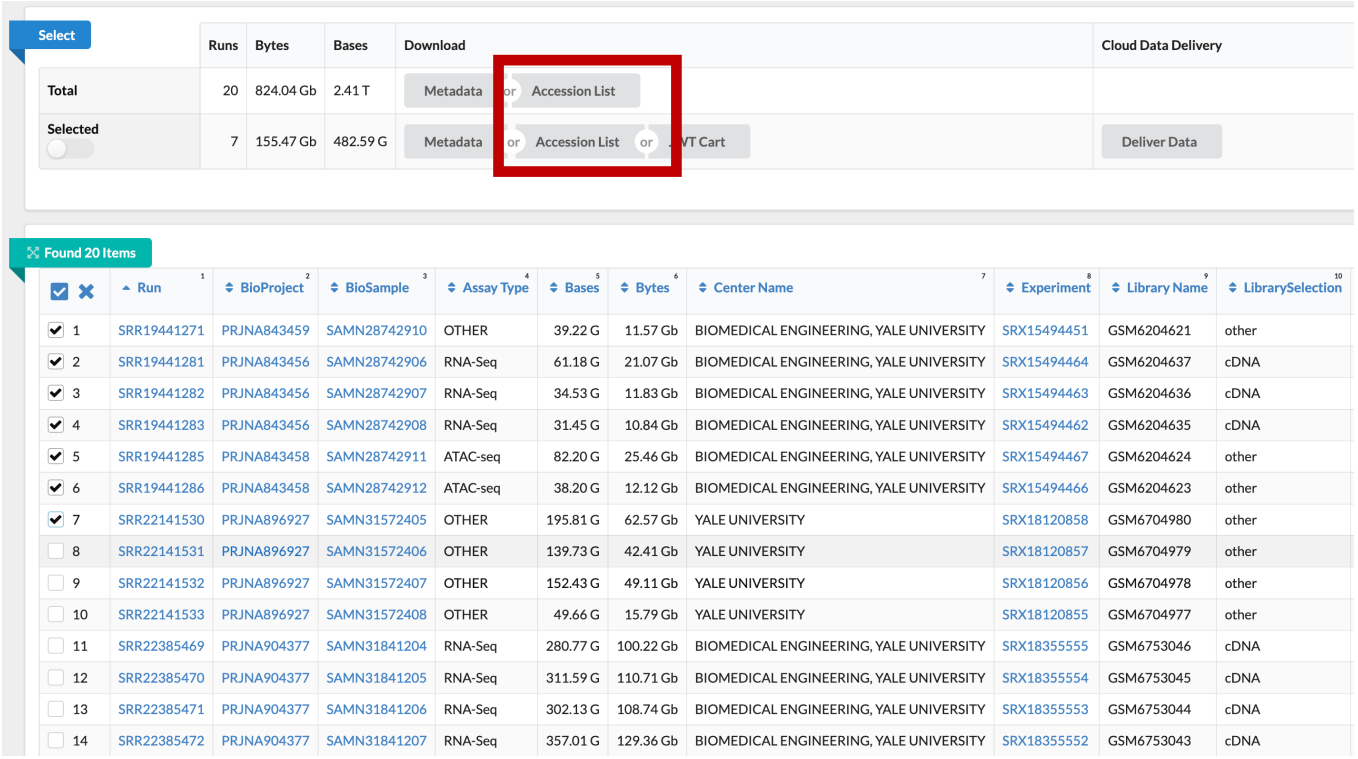
NCBI-SRA

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



NCBI-SRA

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。



Select

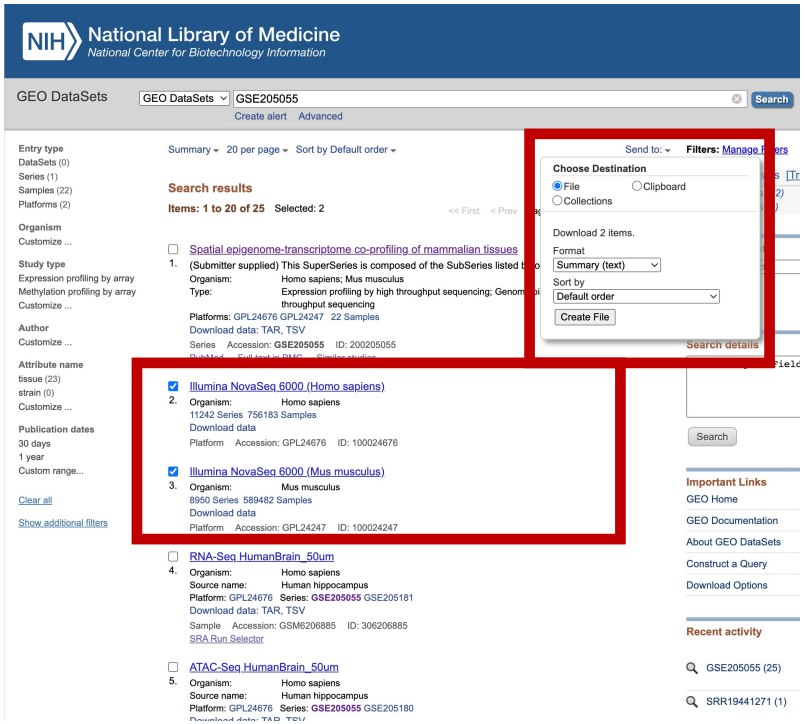
	Runs	Bytes	Bases	Download	Cloud Data Delivery
Total	20	824.04 Gb	2.41 T	Metadata or Accession List	
Selected	7	155.47 Gb	482.59 G	Metadata or Accession List or VT Cart	Deliver Data

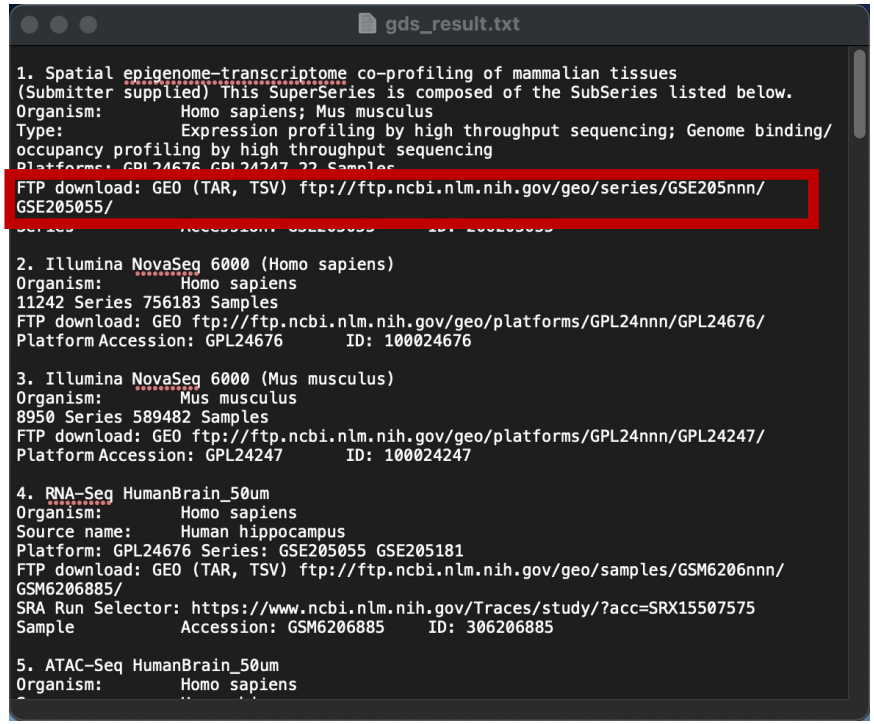
Found 20 Items

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

NCBI-SRA

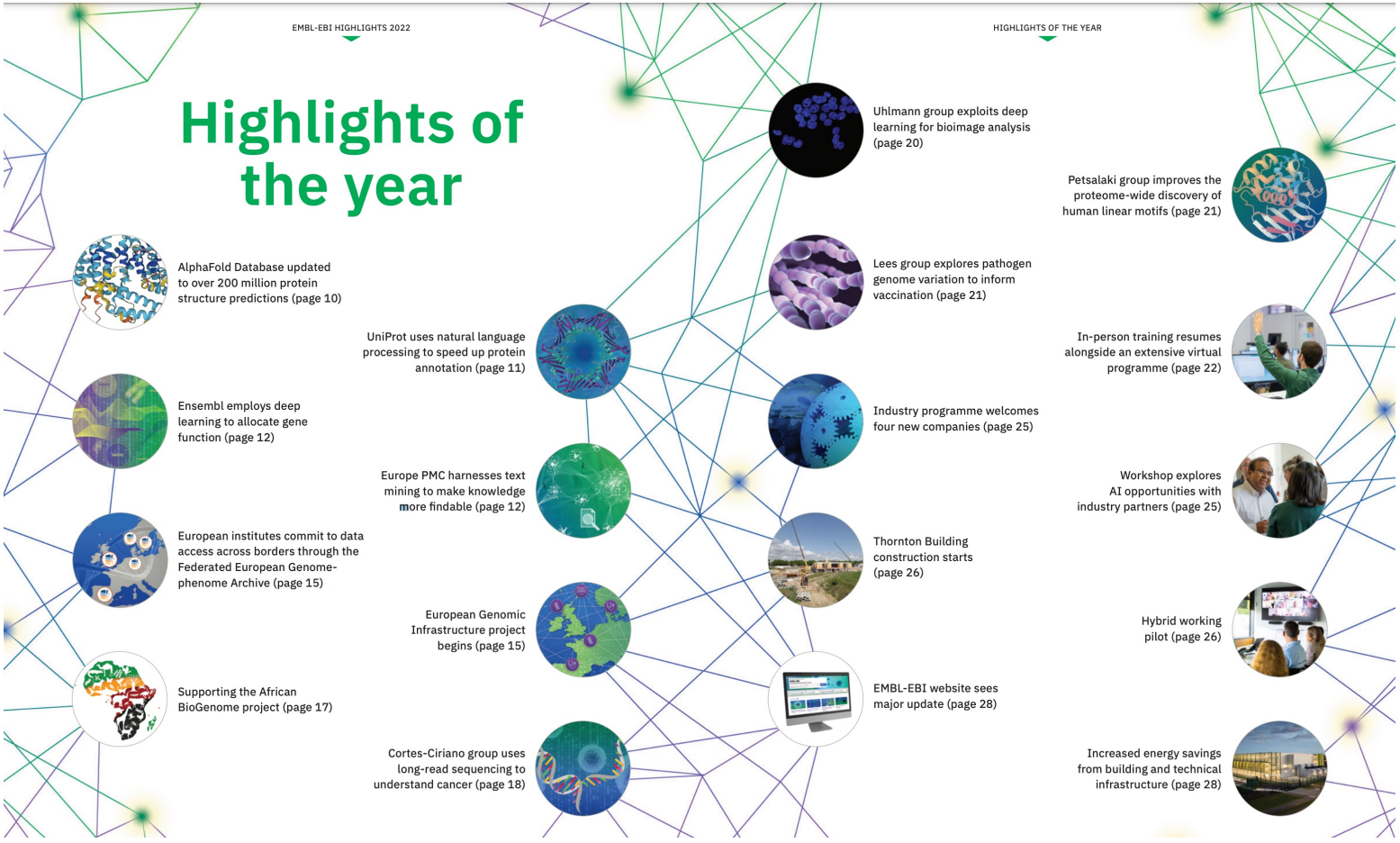
或者在GEO数据库中搜索GSE号，勾选对应的样本，导出summary file。找到对应的ftp下载地址，使用wget或者curl命令下载。





EMBL-EBI

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。



EMBL-EBI

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等基因变到相关的数据库。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

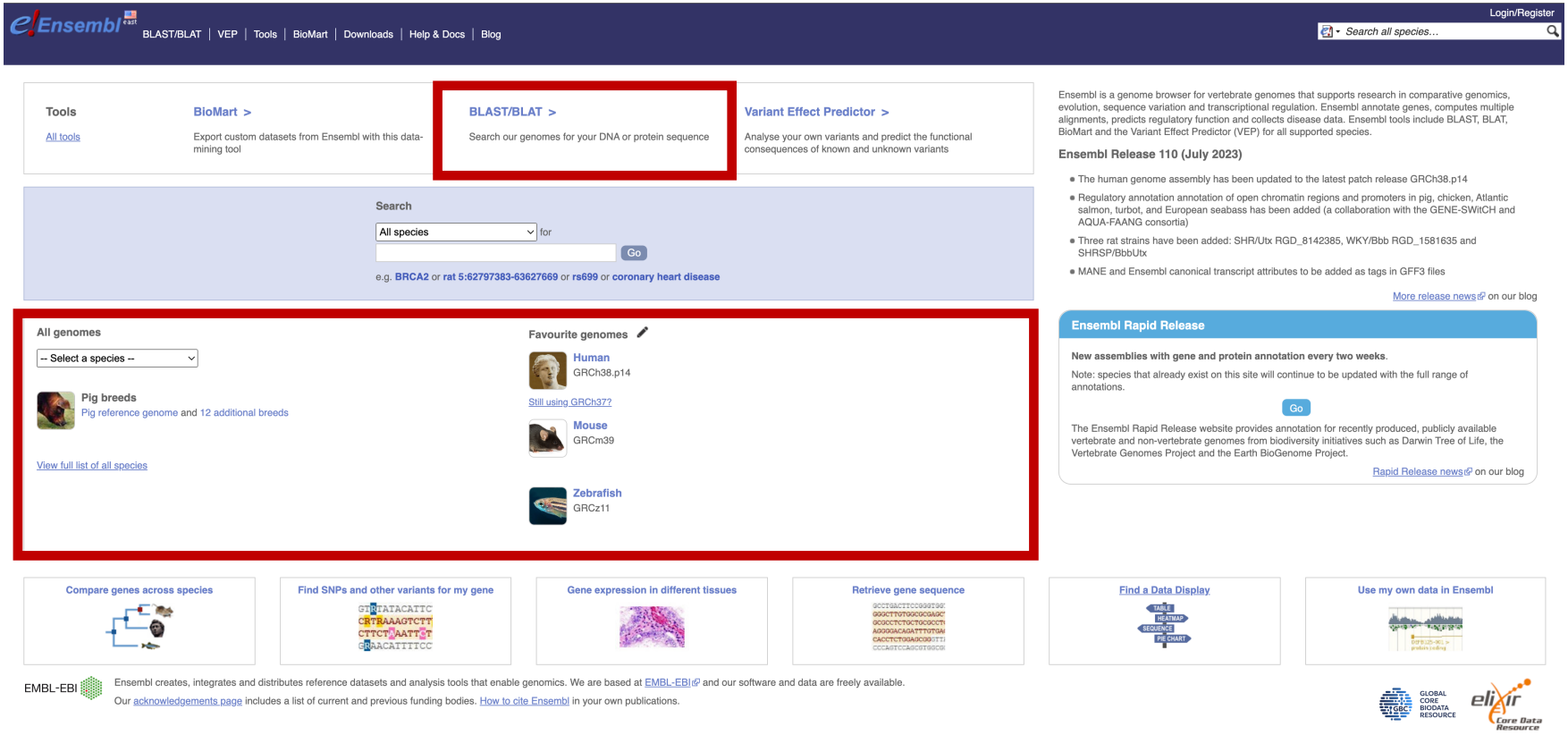
Annotation Platform

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**。



EMBL-EBI Ensembl

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

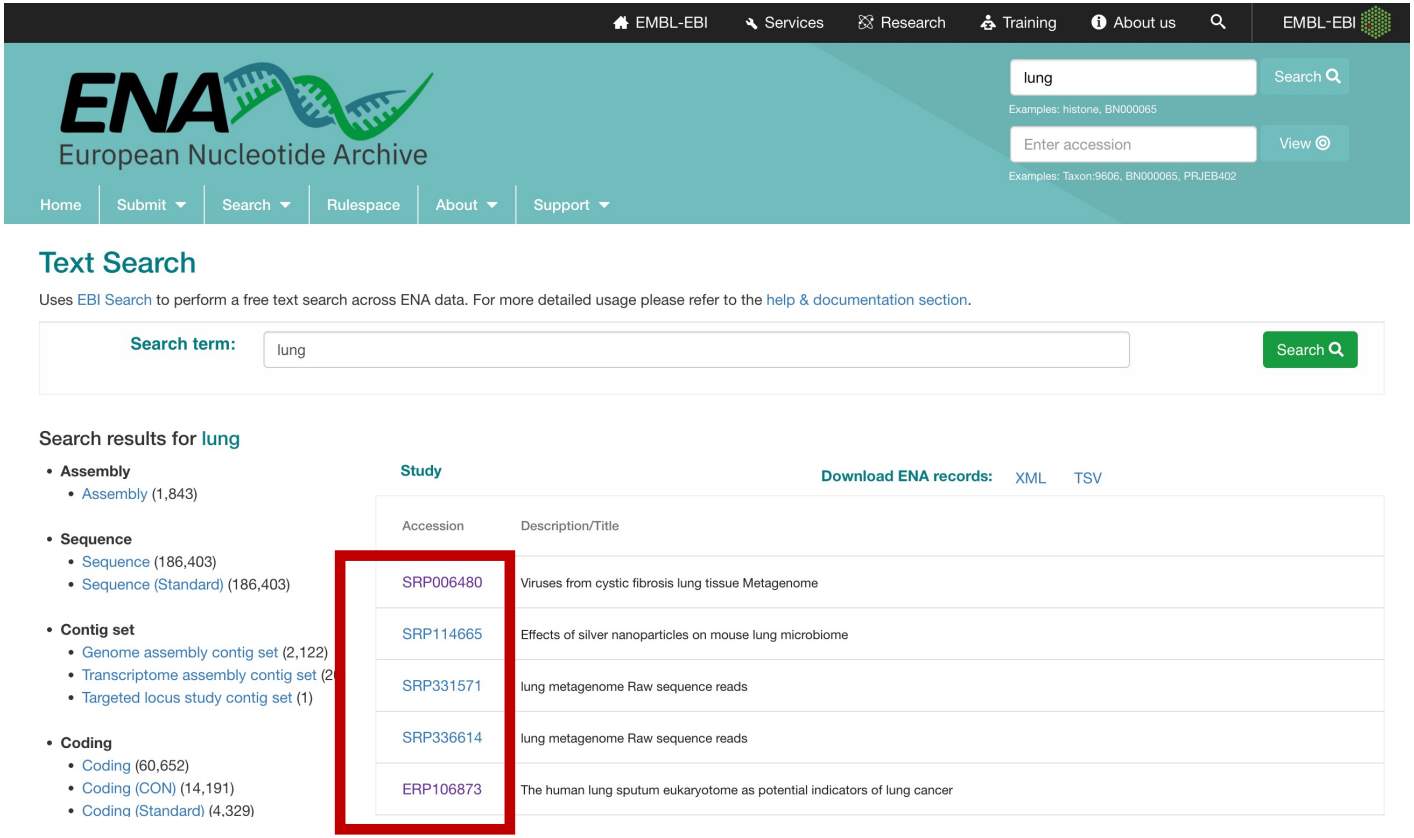
Assembly	GRCh38.p14 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.29, Dec 2013
Base Pairs	3,099,750,718
Golden Path Length	3,099,750,718

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

EMBL-EBI ENA

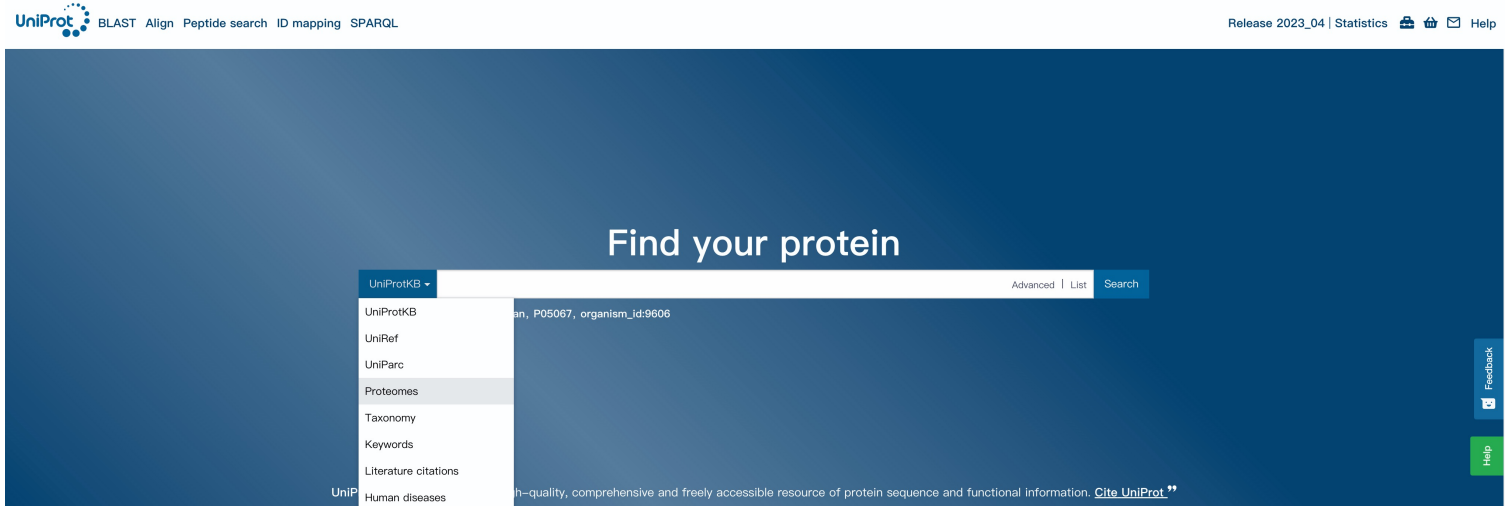
在科研人员完成项目时，也会有科研人员将项目相关的数据上传到EBI的数据库，在ENA数据库中也能够直接搜索。通过不同数据库上传的Study Accession号会有区别。

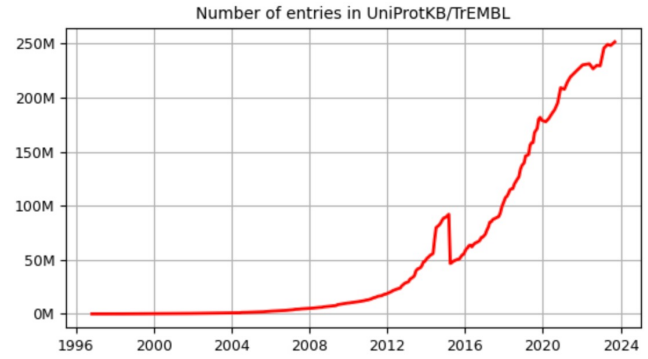


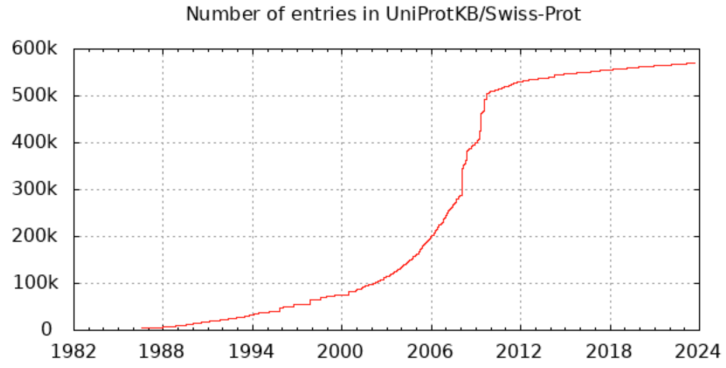
EMBL-EBI UniProt

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。

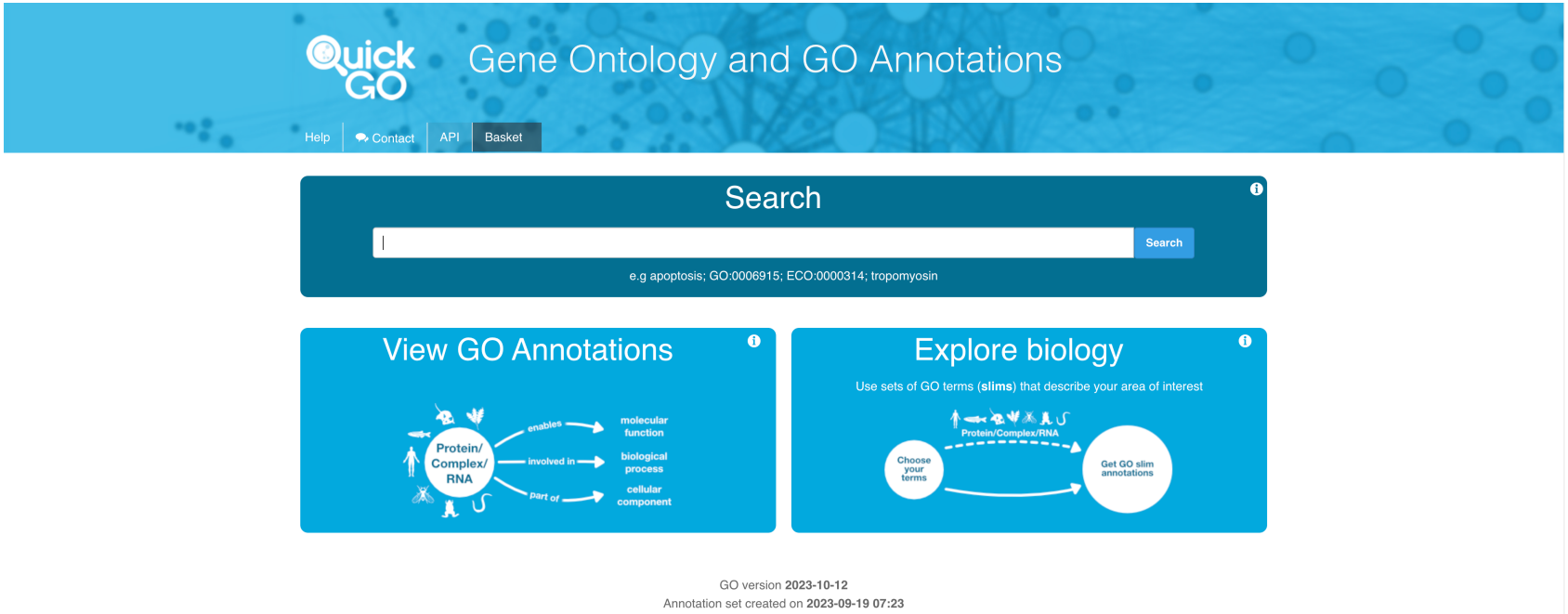






EMBL-EBI Quick GO

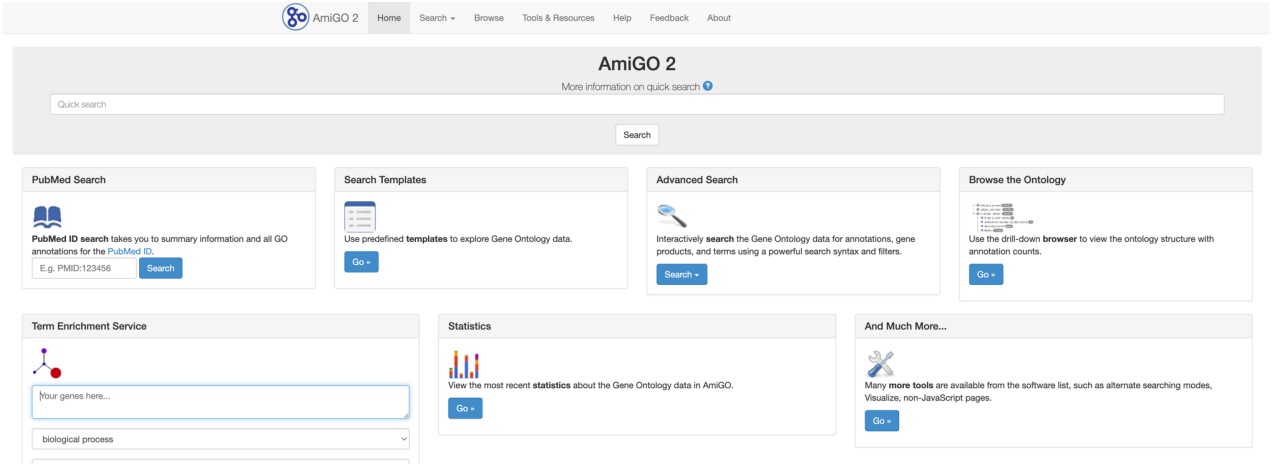
GO（Gene Ontology）是一个标准化系统，用于描述跨物种的基因和基因产物属性，促进生物功能、过程和细胞成分的一致注释。QuickGO 简化并加速了访问和使用 GO 术语的过程.



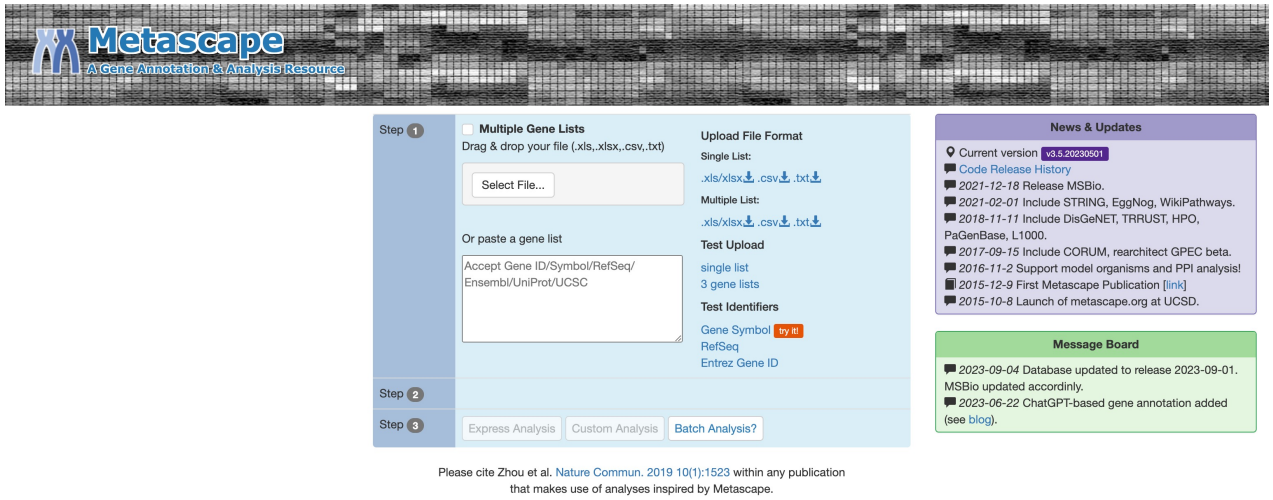
网页在线基因注释工具 - Ami GO & Metascape

将感兴趣的基因list输入，就可以得到基因注释的结果。

AmiGO

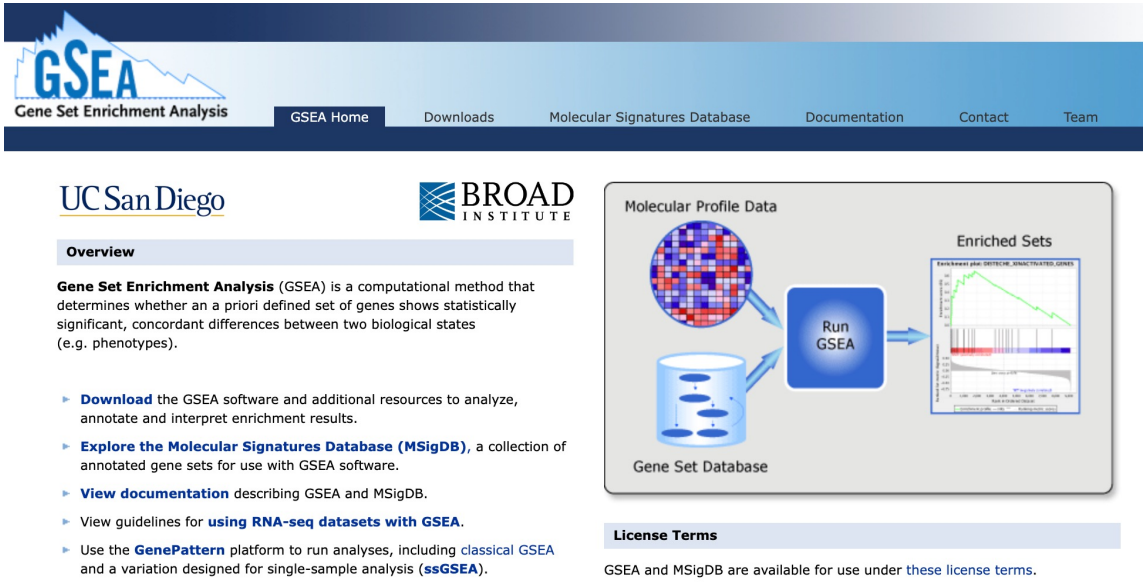


Metascape



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 被引用 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 被引用 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。

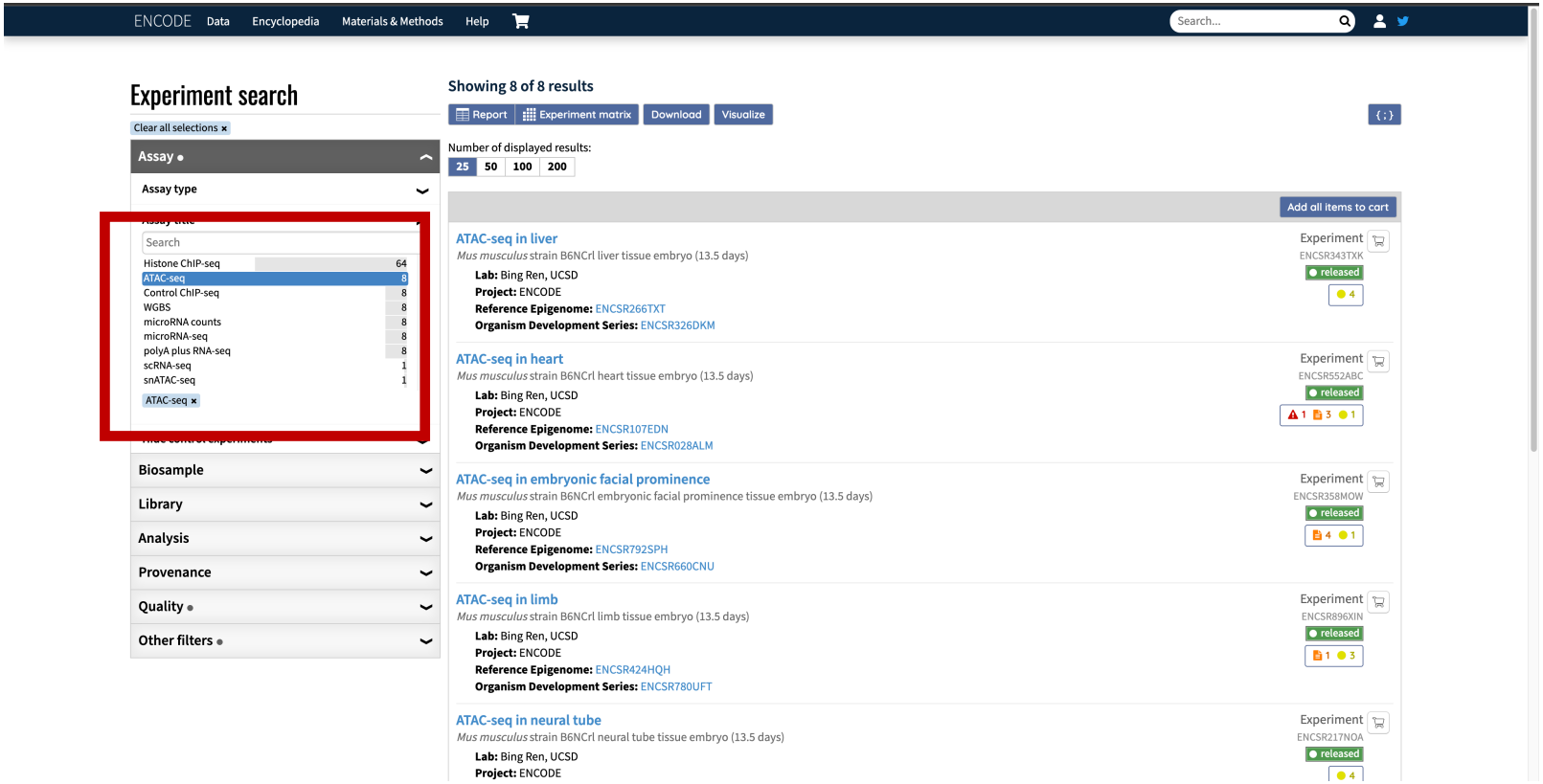
Data availability

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.

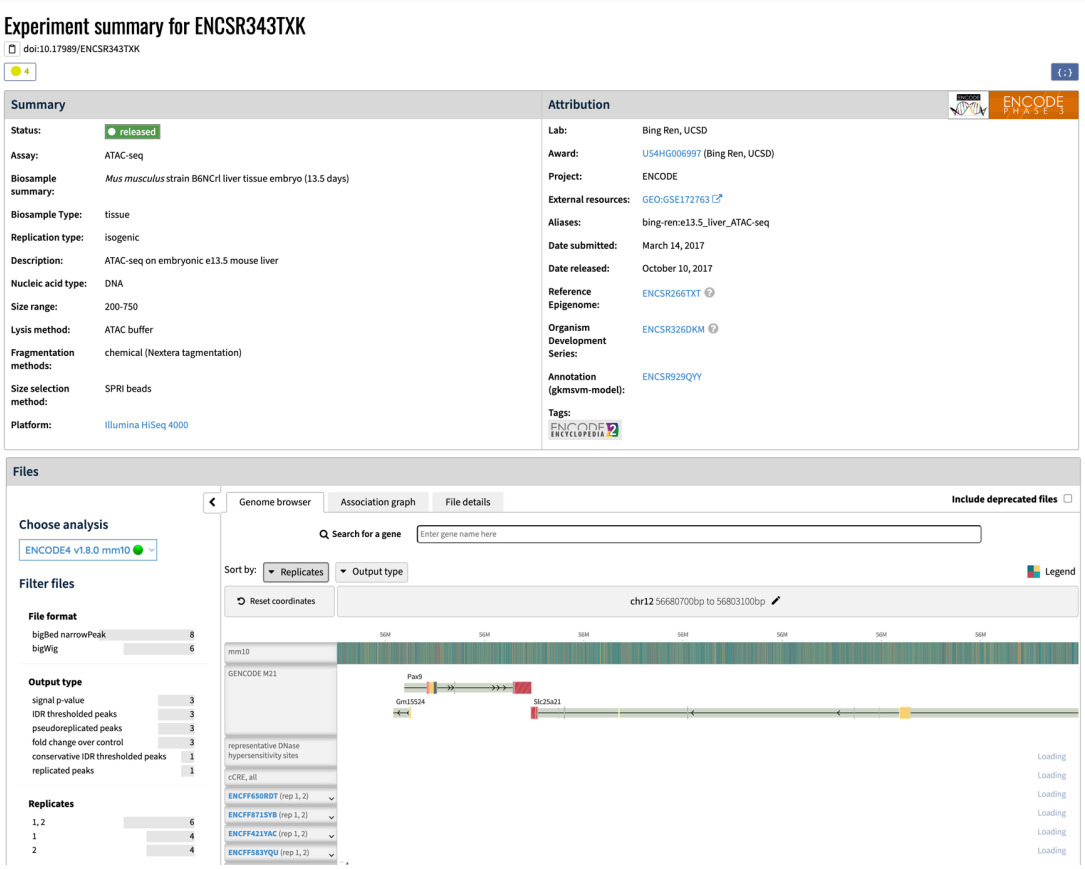
ENCODE

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。



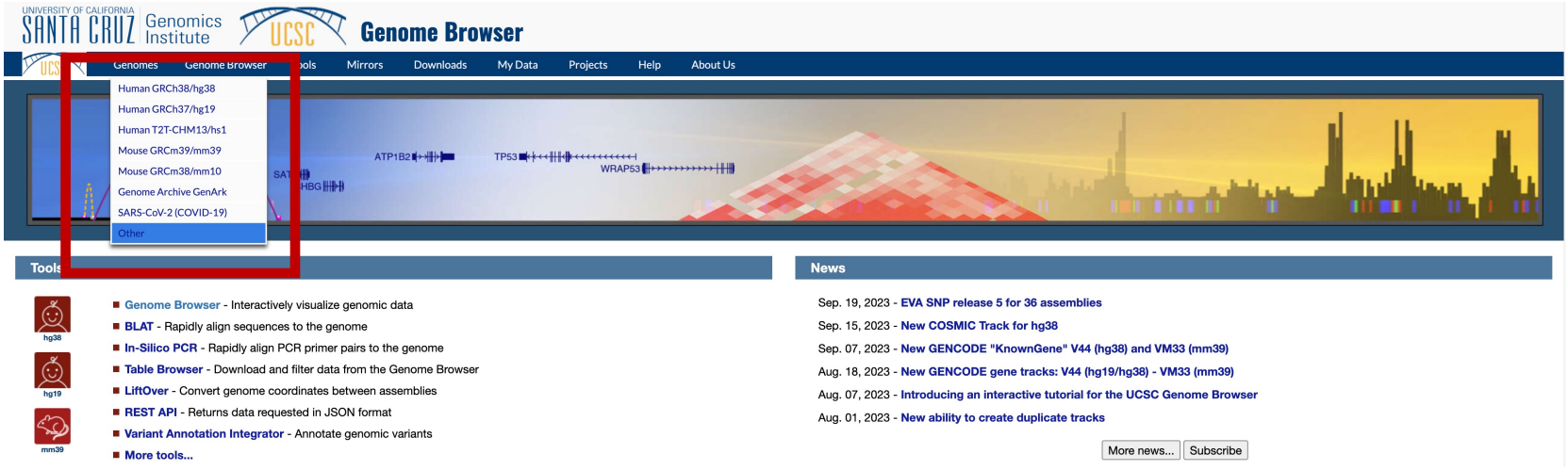
ENCODE

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。



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。



UCSC & WashU Genome Browser

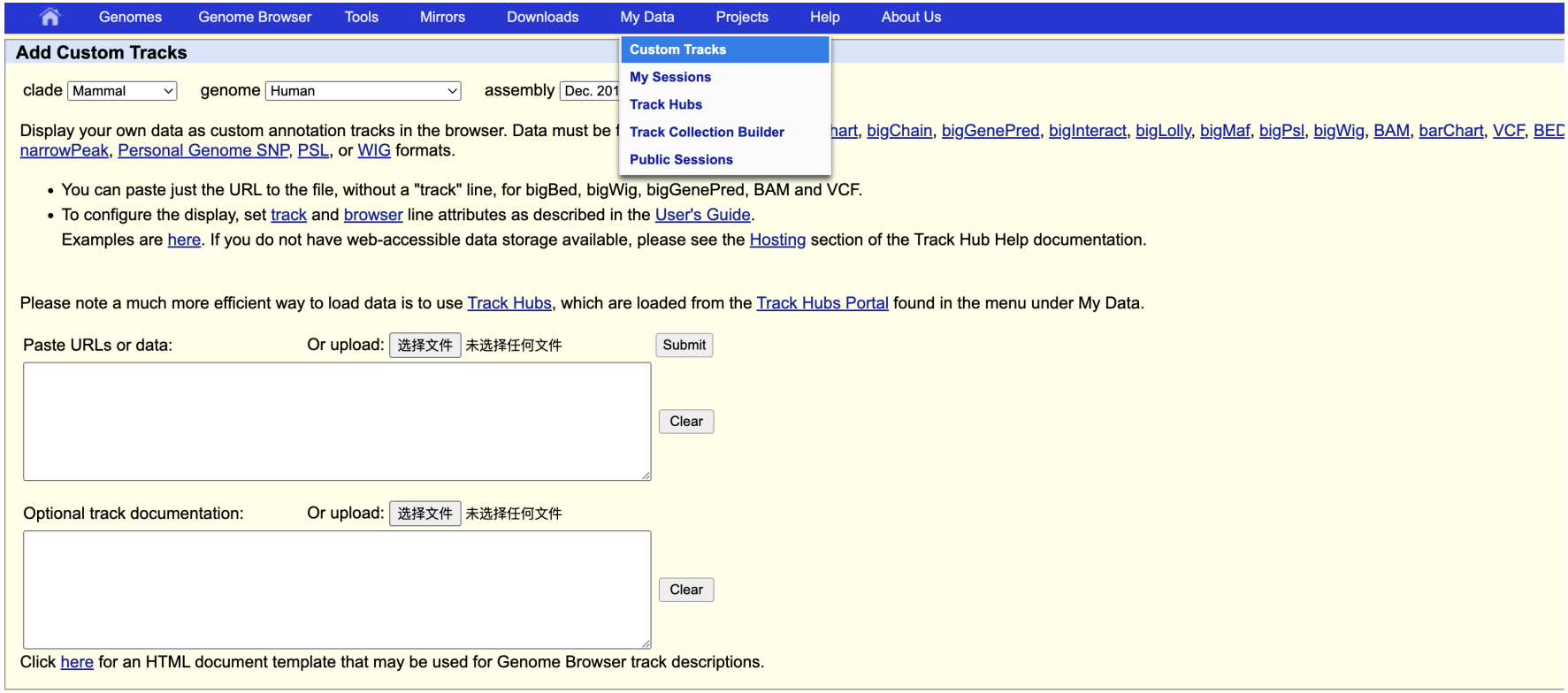
在UCSC genome browser中，可以直接搜索基因或者染色质位置定位，能够看到数据库对应的数据的可视化结果。

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。

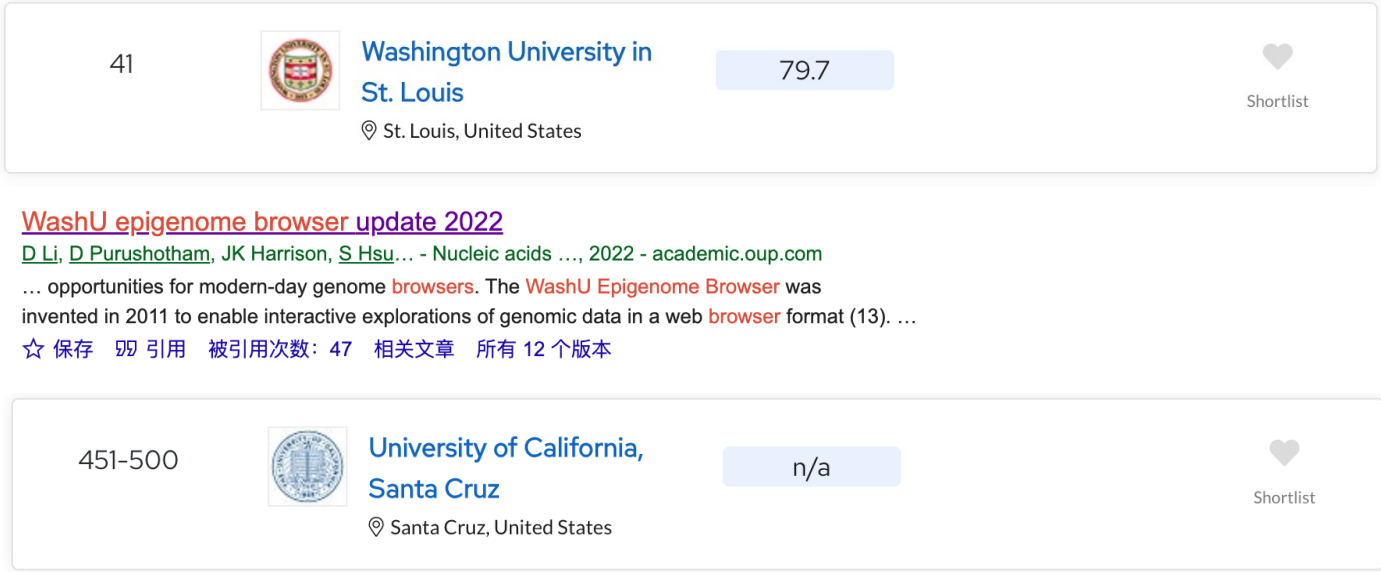
UCSC & WashU Genome Browser

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。



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。



UCSC & WashU Genome Browser

Published: 29 November 2011

The Human Epigenome Browser at Washington University

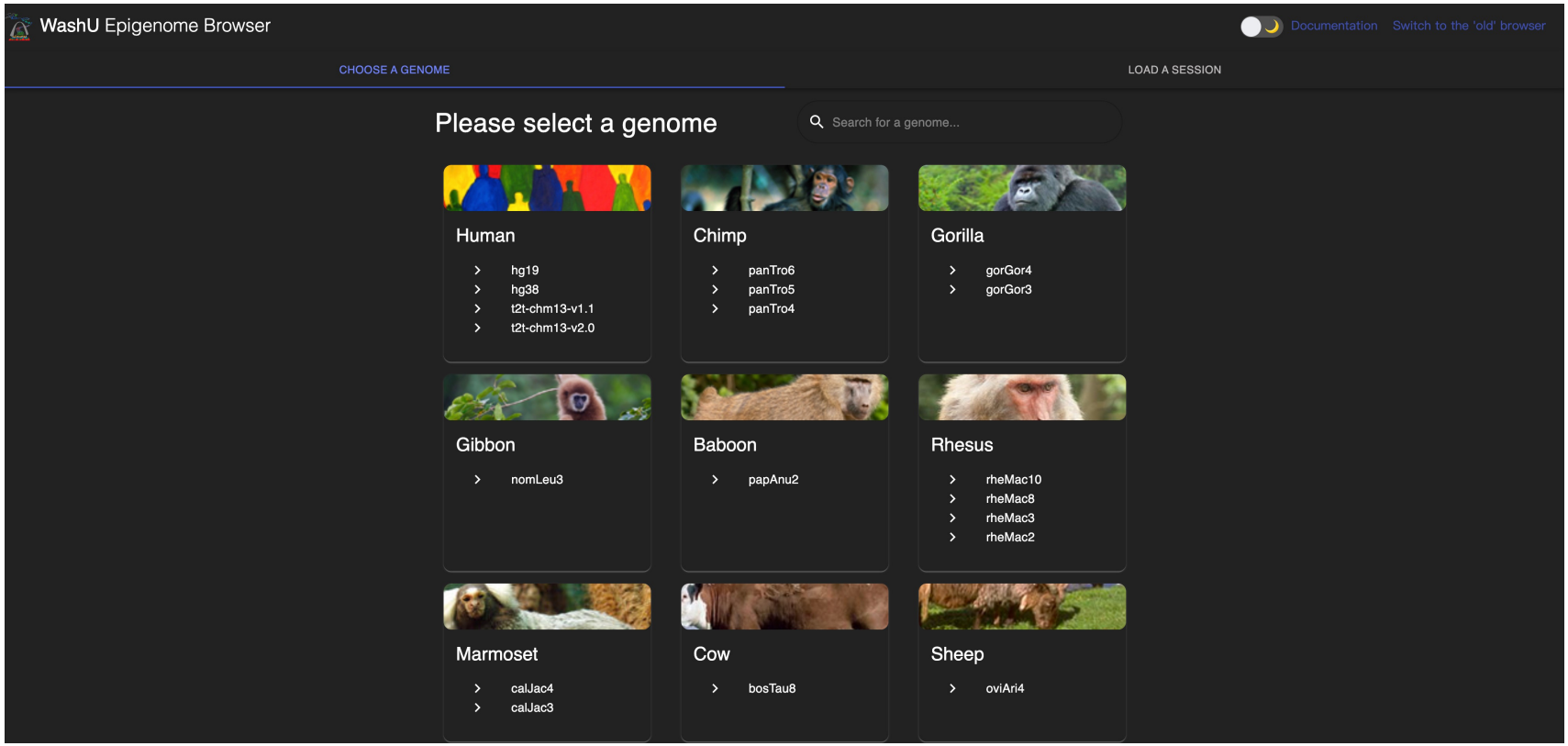
Xin Zhou, Brett Maricque, Mingchao Xie, Daofeng Li, Vasavi Sundaram, Eric A Martin, Brian C Koebbe, Cydney Nielsen, Martin Hirst, Peggy Farnham, Robert M Kuhn, Jingchun Zhu, Ivan Smirnov, W James Kent, David Haussler, Pamela A F Madden, Joseph F Costello ✉ & Ting Wang ✉

Nature Methods **8**, 989–990 (2011) | Cite this article

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

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。



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。