BerryGenomics
贝瑞基因

NCBI数据库

王鹏

科技服务技术支持部

0. NCBI intro

Understanding nature's mute but elegant language of living cells is the quest of modern molecular biology.

Claude Pepper

National Institutes of Health > National Library of Medicine

National Center for Biotechnology Information

开发新的信息技术，助力探索基础分子生物学和遗传过程，从而控制健康和疾病

- 开发自动化系统存储和分析各种数据【分子生物学、生物化学，和遗传学】
- 收集整理国家和国际相关单位的生物技术信息
- 研究先进的计算生物学信息方法，用来分析重要生物学分子的结构和功能
- 促进研究者和医学机构对数据库和软件的使用

0. NCBI database

Literature	
Bookshelf	895,668
MeSH	348,659
NLM Catalog	1,632,085
PubMed	33,300,901
PubMed Central	7,498,879

Genes	
Gene	35,065,002
GEO DataSets	4,924,111
GEO Profiles	128,414,055
HomoloGene	141,268
PopSet	369,822

Proteins	
Conserved Domains	62,852
Identical Protein Groups	461,863,014
Protein	987,174,127
Protein Family Models	179,503
Structure	184,205

Genomes	
Assembly	1,143,972
BioCollections	8,501
BioProject	550,161
BioSample	21,141,887
Genome	66,362
Nucleotide	482,370,507
SRA	17,859,082
Taxonomy	0

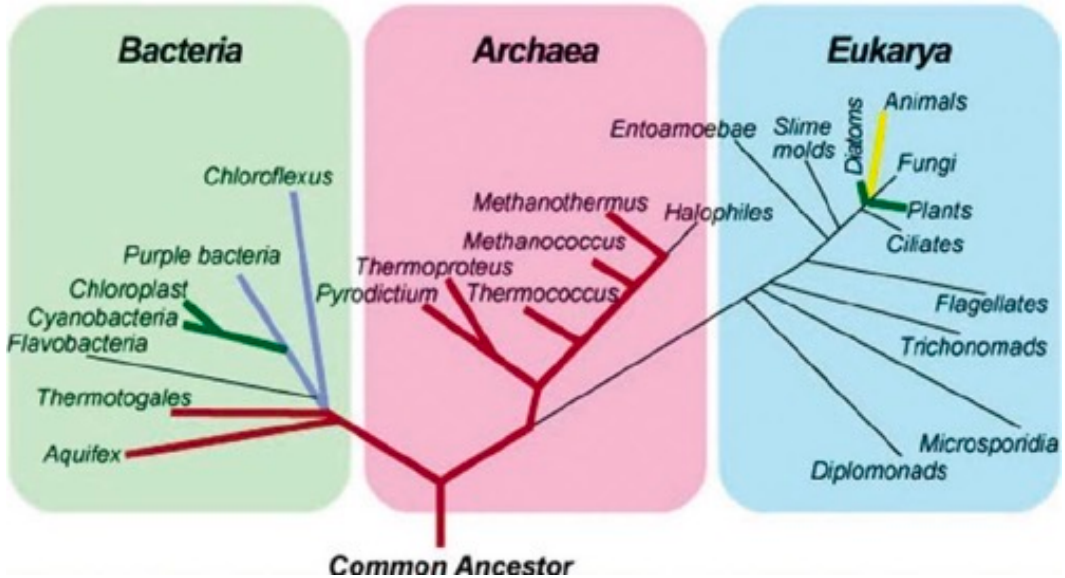
Clinical	
ClinicalTrials.gov	382
ClinVar	1,180,336
dbGaP	1,405
dbSNP	1,076,992,604
dbVar	7,193,811
GTR	78,535
MedGen	204,742
OMIM	27,474

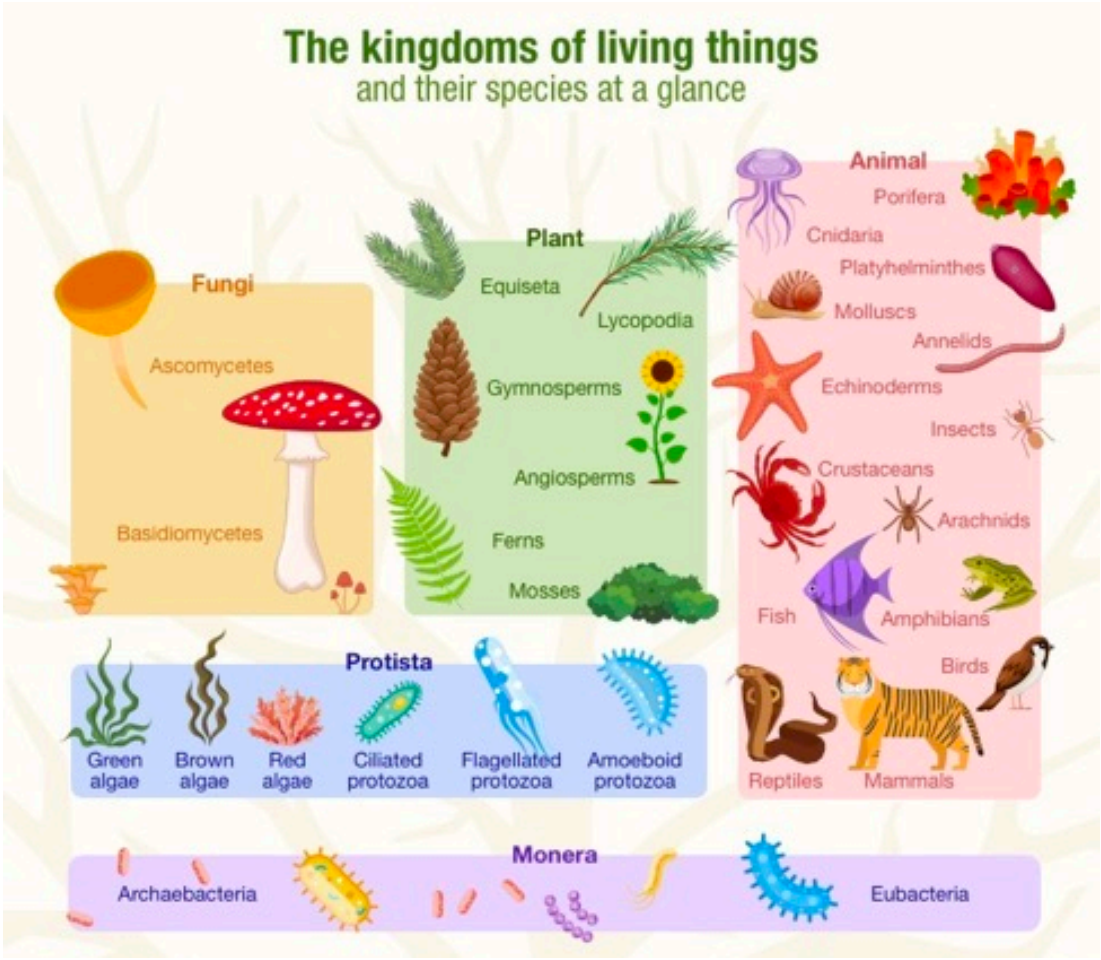
PubChem	
BioAssays	0
Compounds	0
Pathways	0
Substances	0

6大类；35个database

https://www.ncbi.nlm.nih.gov/search/all/?term=all[filter]

1. Taxonomy

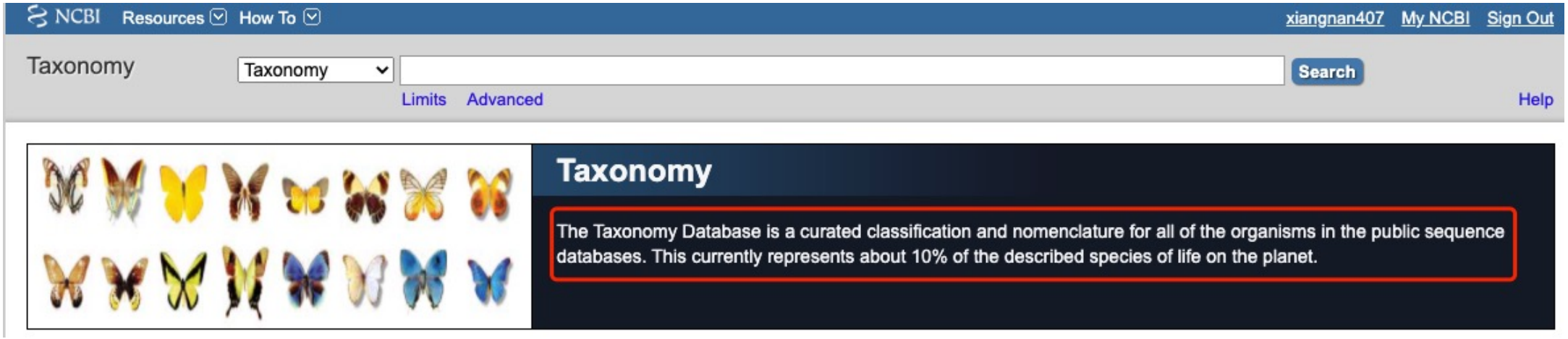


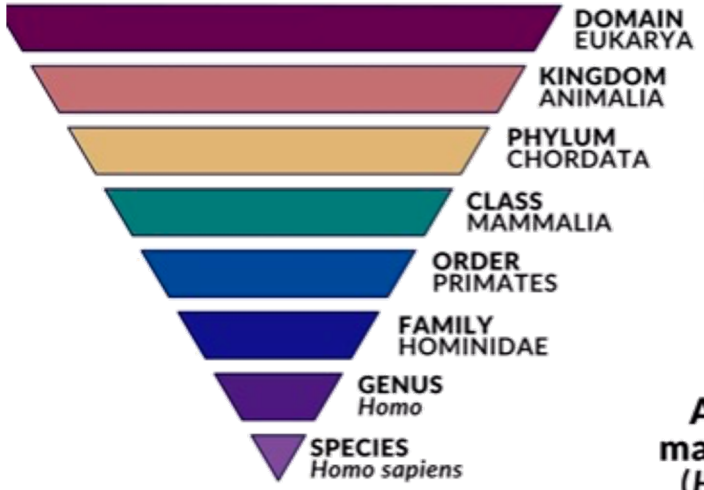


Three domains of life; Five kingdoms of life

https://www.iberdrola.com/sustainability/biology-kingdoms-living-things-classification

1. Taxonomy





Adult human male and female (*Homo sapiens*)

Contains the names and phylogenetic lineages of more than **160,000 organisms** that have molecular data in the NCBI databases. New taxa are added to the Taxonomy database as data are deposited for them.

https://www.ncbi.nlm.nih.gov/taxonomy

1. Taxonomy

https://ftp.ncbi.nih.gov/pub/taxonomy/

names.dmp. nodes.dmp.

https://ftp.ncbi.nih.gov/pub/taxonomy/taxdump_readme.txt

将names和nodes转成界门纲目科属种

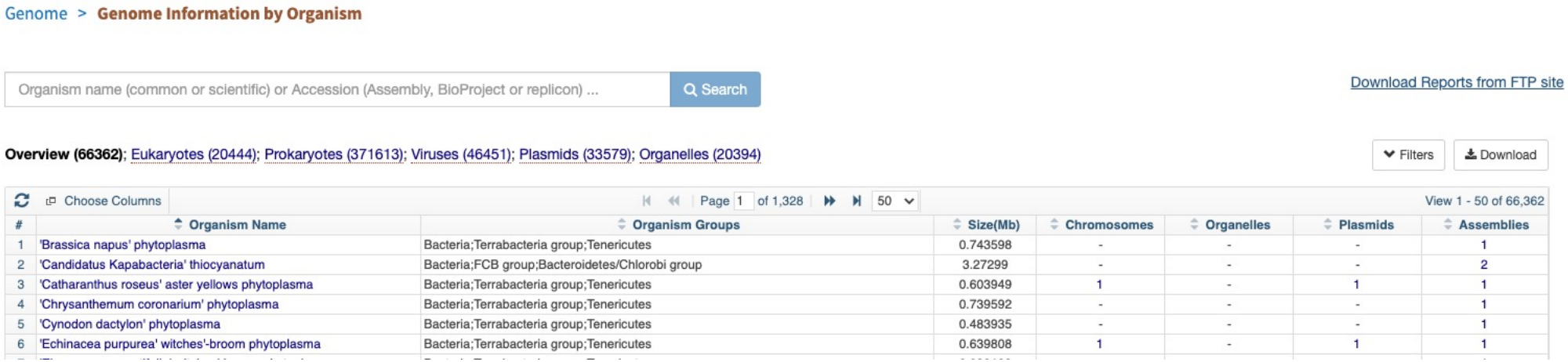
https://github.com/wangpeng407/ncbi_taxa

k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Hyphomicrobiales;f__Xanthobacteraceae;g__Azorhizobium;s__Azorhizobium caulinodans

k__Eukaryota;p__Chordata;c__Mammalia;o__Primates;f__Hominidae;g__Homo;s__Homo sapiens

2. Genome

https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/



Genome > **Genome Information by Organism**

Organism name (common or scientific) or Accession (Assembly, BioProject or replicon) ... Search

Download Reports from FTP site

Overview (66362); Eukaryotes (20444); Prokaryotes (371613); Viruses (46451); Plasmids (33579); Organelles (20394)

Filters Download

Choose Columns | Page 1 of 1,328 | 50 | View 1 - 50 of 66,362

#	Organism Name	Organism Groups	Size(Mb)	Chromosomes	Organelles	Plasmids	Assemblies
1	'Brassica napus' phytoplasma	Bacteria;Terrabacteria group;Tenericutes	0.743598	-	-	-	1
2	'Candidatus Kapabacteria' thiocyanatum	Bacteria;FCB group;Bacteroidetes/Chlorobi group	3.27299	-	-	-	2
3	'Catharanthus roseus' aster yellows phytoplasma	Bacteria;Terrabacteria group;Tenericutes	0.603949	1	-	1	1
4	'Chrysanthemum coronarium' phytoplasma	Bacteria;Terrabacteria group;Tenericutes	0.739592	-	-	-	1
5	'Cynodon dactylon' phytoplasma	Bacteria;Terrabacteria group;Tenericutes	0.483935	-	-	-	1
6	'Echinacea purpurea' witches'-broom phytoplasma	Bacteria;Terrabacteria group;Tenericutes	0.639808	1	-	1	1

https://ftp.ncbi.nlm.nih.gov/genomes/

https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/

3. GenBank

International Nucleotide Sequence Database Collaboration (INSDC)

GeneBank　ENA　DDBJ

三大数据库，每日互换数据

- 每个机构或研究者提交的、可公开的核酸序列
- 这些序列必须是注释好的，有相应的cds序列
- 数据库中有大量的冗余序列
- 每条序列的记录归初始提交者所有，第三方不可更改

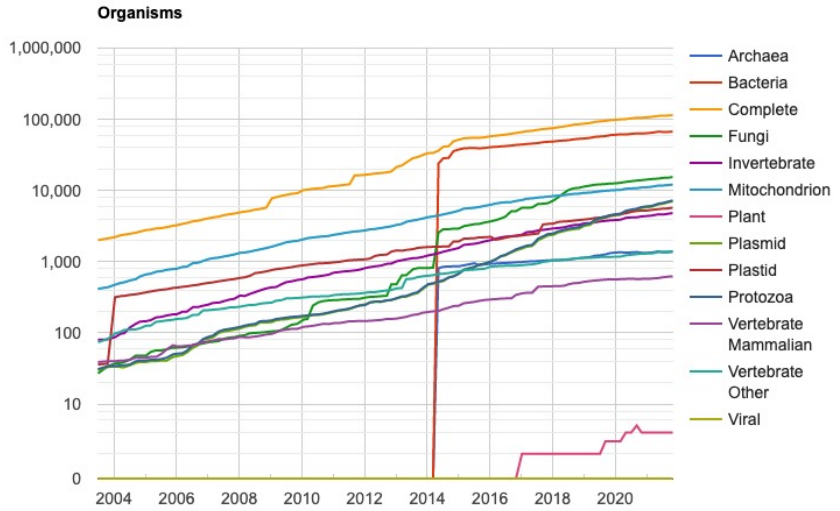
- Single gene or mRNA sequences
- Complete genomes
- Whole Genome Shotgun (WGS)
- Transcriptome Shotgun Assembly (TSA)
- Third-Party annotation (TPA)
- Expressed Sequence Tags (EST)
- Targeted Locus Study (TLS)

https://ftp.ncbi.nlm.nih.gov/genbank/

4. RefSeq

NCBI reference sequence (RefSeq)

- A comprehensive, integrated, non-redundant, well-annotated set of sequences
- Including genomic DNA, transcripts, and proteins
- Providing a stable reference for medical, functional, and diversity studies
- RefSeq genomes are copies of selected assembled genomes available in GenBank
- Transcripts and proteins：submitter’s annotation, curated annotation from model organism, NCBI annotation



Announcements

November 5, 2021
RefSeq Release 209 is available for FTP

This release includes:

Proteins: 215,655,378
Transcripts: 41,751,205
Organisms: 114,396
Available at: ftp://ftp.ncbi.nlm.nih.gov/refseq/release/
Documentation: Release Notes

See previous announcements, follow NCBI on Twitter, or subscribe to NCBI's refseq-announce mail list to receive announcements.

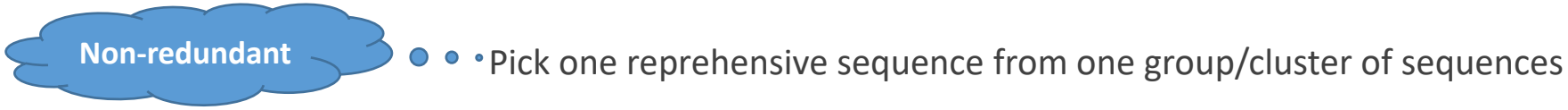
https://ftp.ncbi.nlm.nih.gov/refseq/

5. NR & NT

NT

Nucleotide collection (nt): The nucleotide collection consists of **GenBank+EMBL+DDBJ+PDB+RefSeq** sequences, but **excludes** EST, STS, GSS, WGS, TSA, patent sequences as well as phase 0, 1, and 2 HTGS sequences and sequences longer than 100Mb. The database is **non-redundant**. Identical sequences have been merged into one entry, while preserving the accession, GI, title and taxonomy information for each entry.

Molecule Type: mixed DNA; Update date: 2021/11/14; NO. of sequences: 76,027,954



NR

Nucleotide collection (nr): **All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF** excluding environmental samples from **WGS** projects.

Molecule Type: Protein; Update date: 2021/11/14; NO. of sequences: 439,976,609

https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/
https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/

6. GEO

Gene expression Omnibus

GEO DataSets　　GEO Profiles

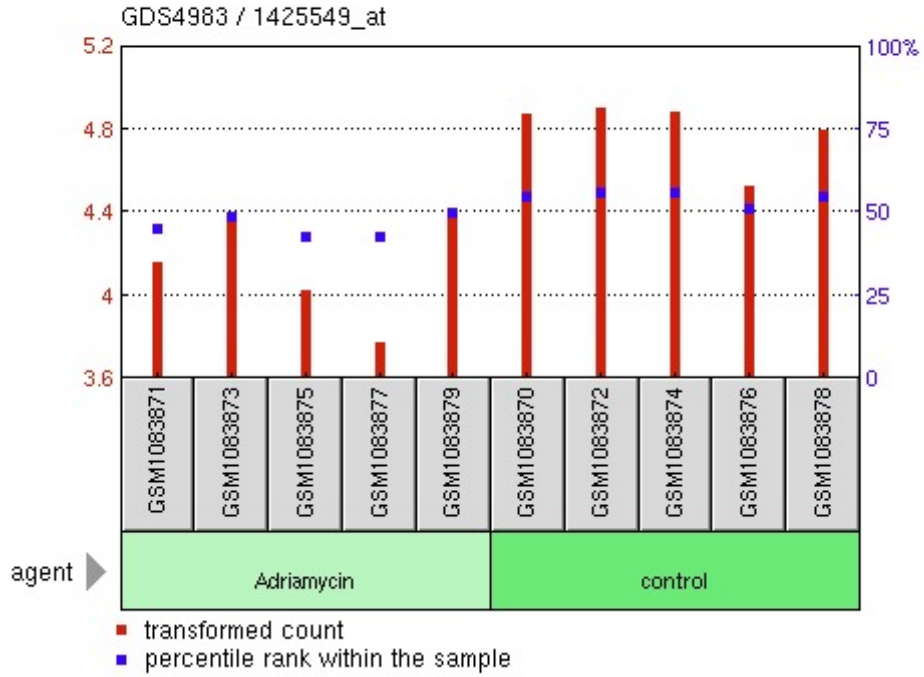
测了什么？ 什么测的？ 测得了什么？

https://www.ncbi.nlm.nih.gov/gds
https://ftp.ncbi.nlm.nih.gov/geo/

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
GSE184019_count_gene_s.txt.gz	518.8 Kb	(ftp)(http)	TXT
GSE184019_tpm_gene_s.txt.gz	670.6 Kb	(ftp)(http)	TXT

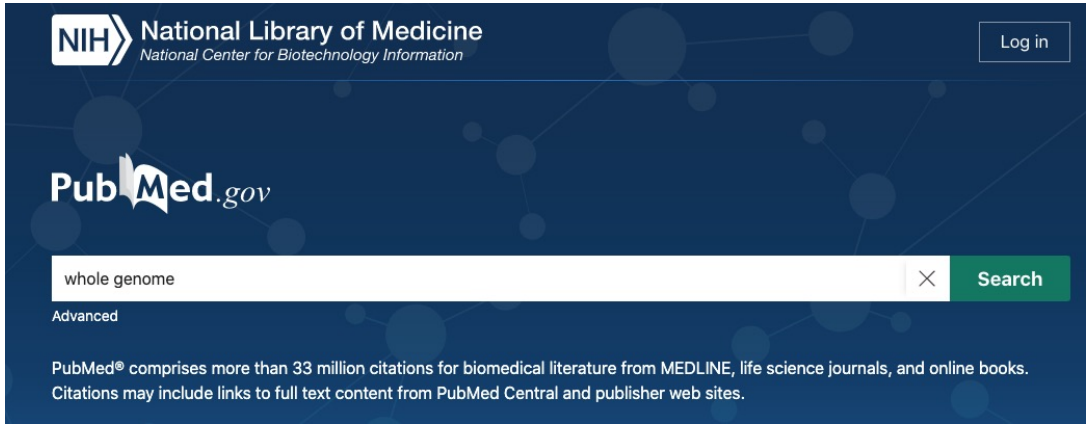
Organism Mus musculus



https://www.ncbi.nlm.nih.gov/geo/geo2r/

以mouse和alcohol为关键词搜索为例　　NCBI中GEP2R可实现对转录组的基本重分析（ GSE25724 ）

7. PubMed & PMC

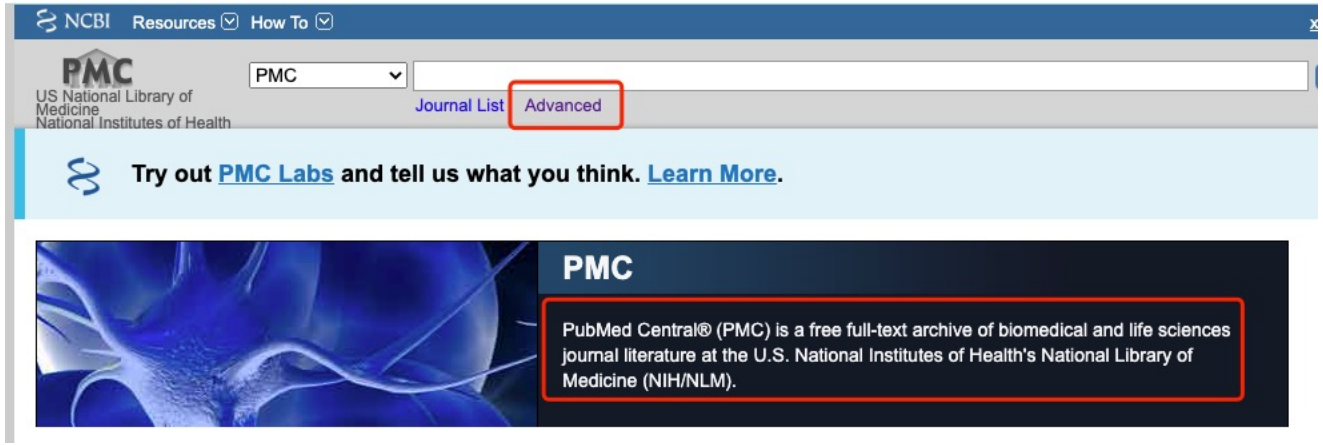


PubMed Central (PMC) ：免费的数字化存档数据库，包含生物医学和生命科学的全文文献。

https://www.ncbi.nlm.nih.gov/pmc/

PubMed：引文数据库，不包含全文但是有>33M的来自其他期刊和书籍的引文。

https://pubmed.ncbi.nlm.nih.gov/



8. NCBI learning materials

Webinars	https://www.ncbi.nlm.nih.gov/home/coursesandwebinars/ https://ftp.ncbi.nlm.nih.gov/pub/education/public_webinars/
Conferences	https://www.ncbi.nlm.nih.gov/home/conferencesandpresentations/ https://ftp.ncbi.nlm.nih.gov/pub/education/Mod_Workshops/
Tutorials	https://www.ncbi.nlm.nih.gov/home/tutorials/
Documentation	https://www.ncbi.nlm.nih.gov/home/documentation/

BerryGenomics
贝瑞基因

Thank You!

官 方 网 站

官 方 微 信

www.berrygenomics.com