National Genomics Data Center

CNCB-NGDC Progress Report

Yiming Bao

Director
National Genomics Data Center
Beijing, China

The 19th ABC Symposium
Dec. 9, 2022 • Zoom

中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

CONTENTS

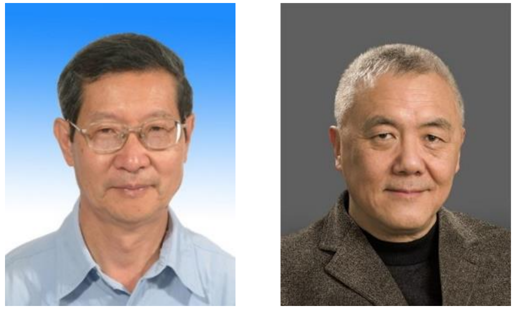
1 Updates

2 Development

3 Future directions

The Team

❑ Steering Advisors

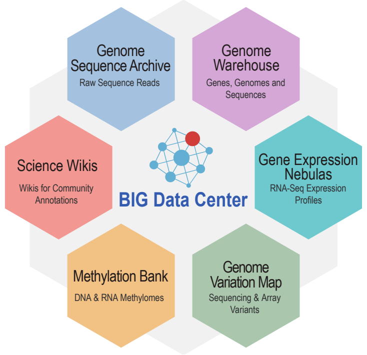


❑ Professors

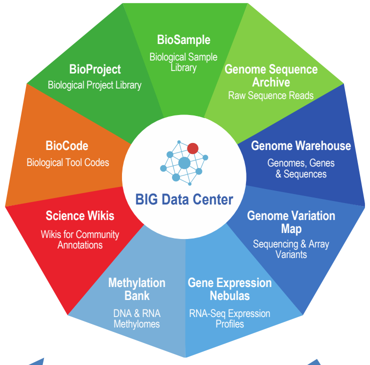
67
students

53
Staff

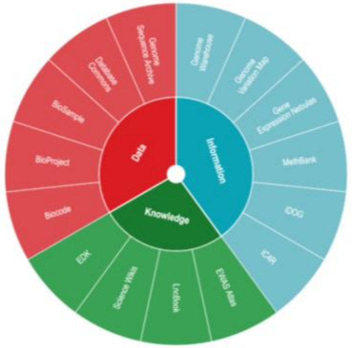
The growing of capability



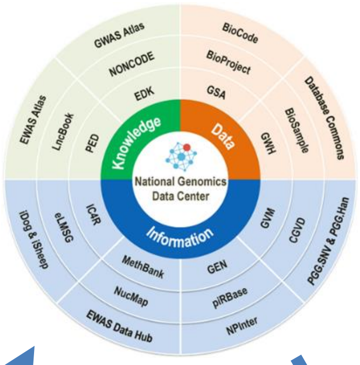
2016



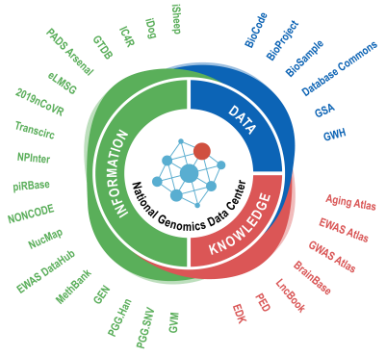
2017



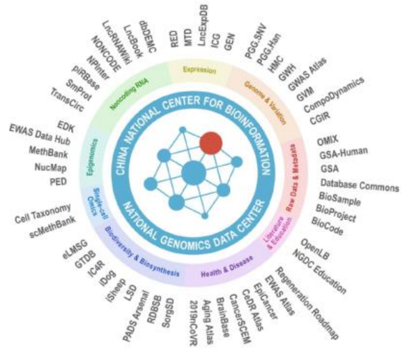
2018



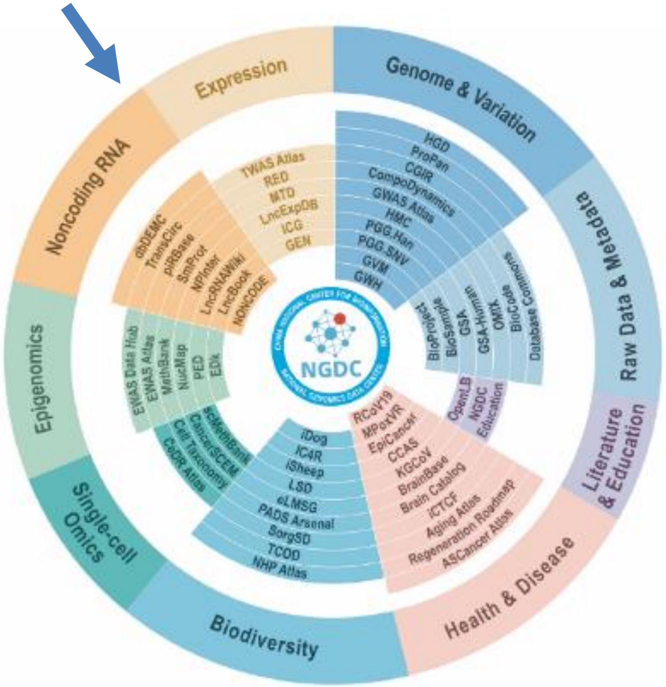
2019



2020

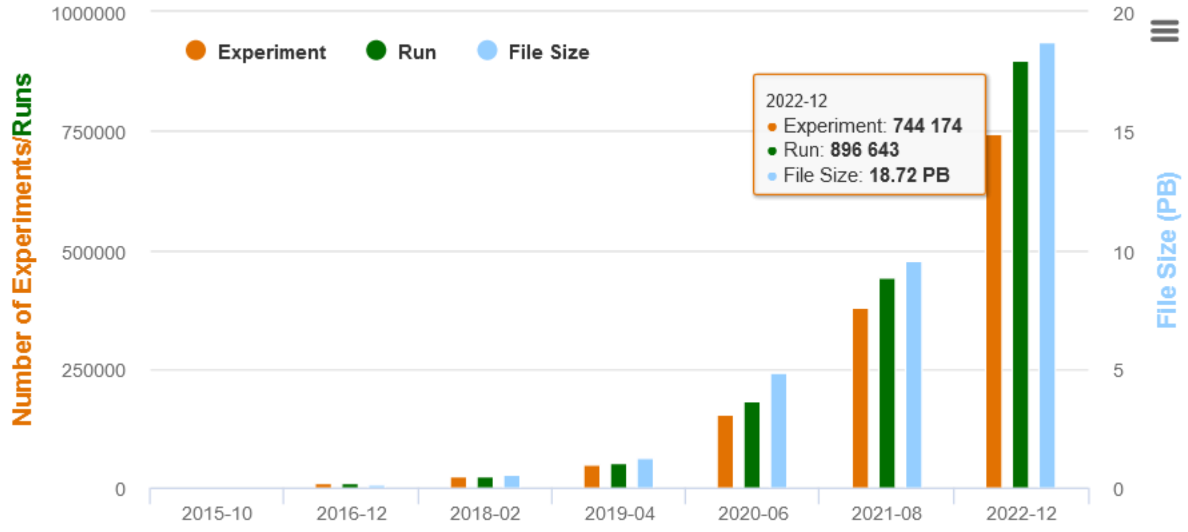


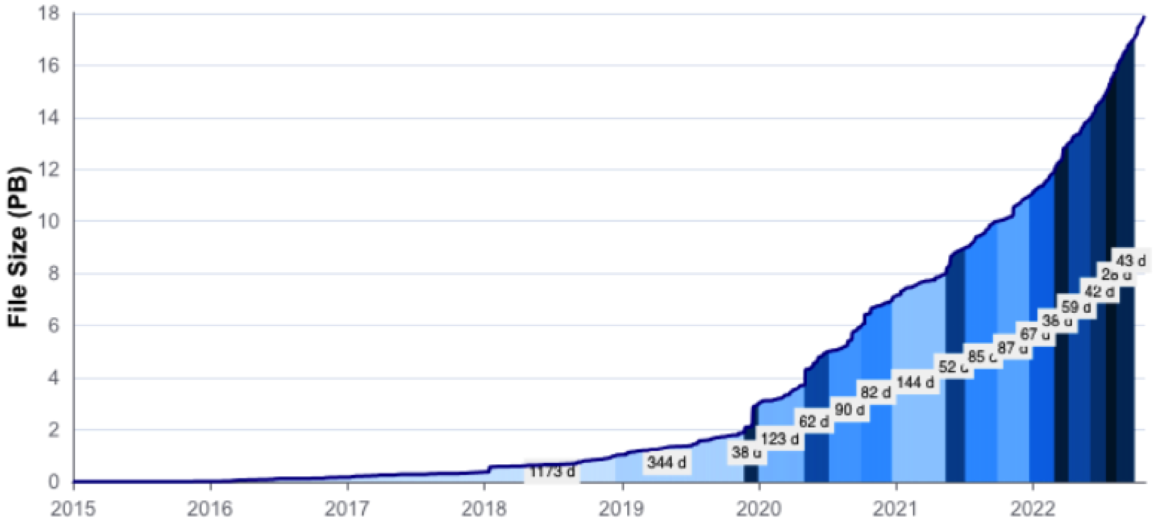
2021



2022

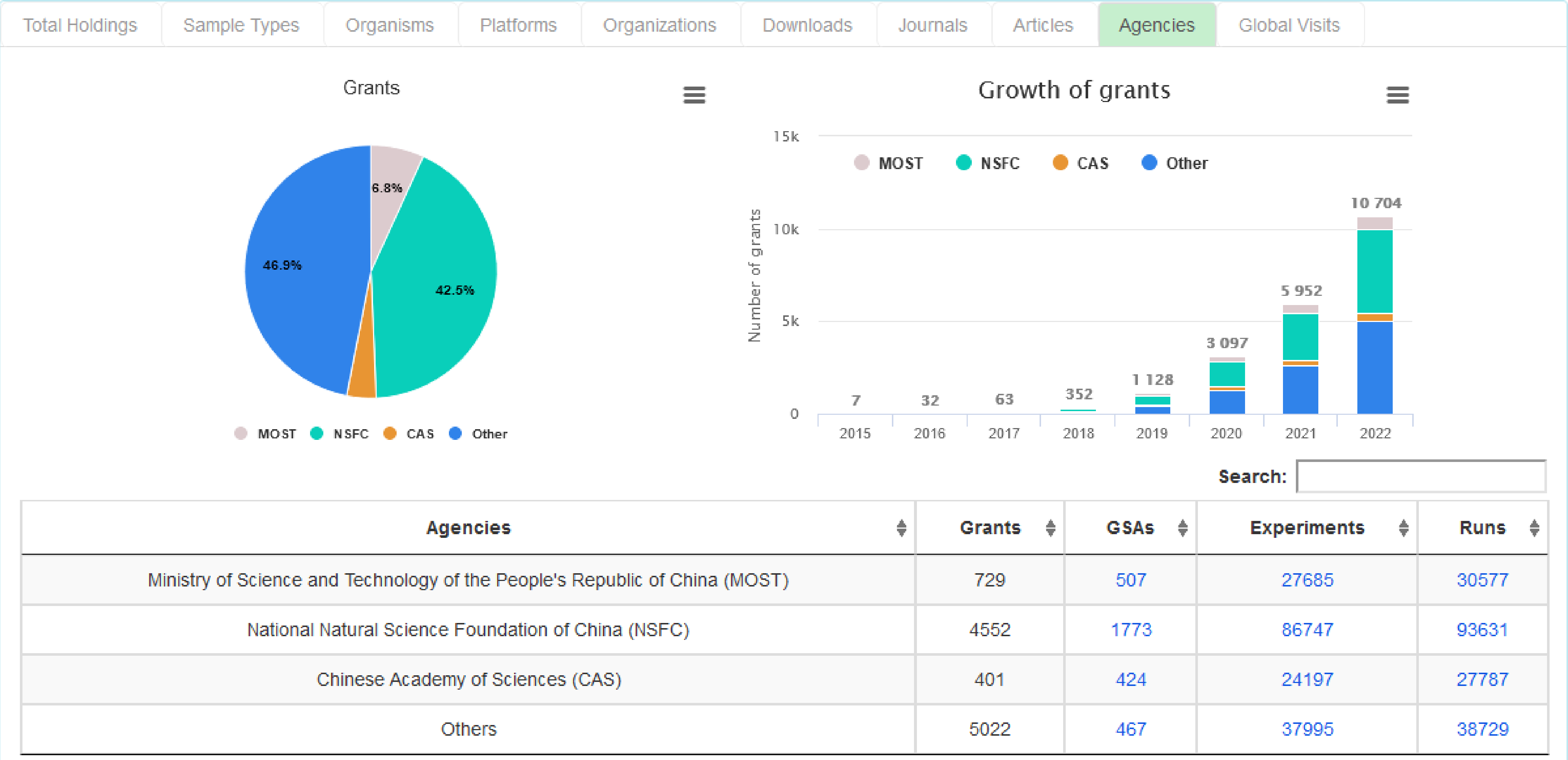
GSA Data Growth





18.7 PB as of 2022-12-09

Grants that support data deposited in GSA



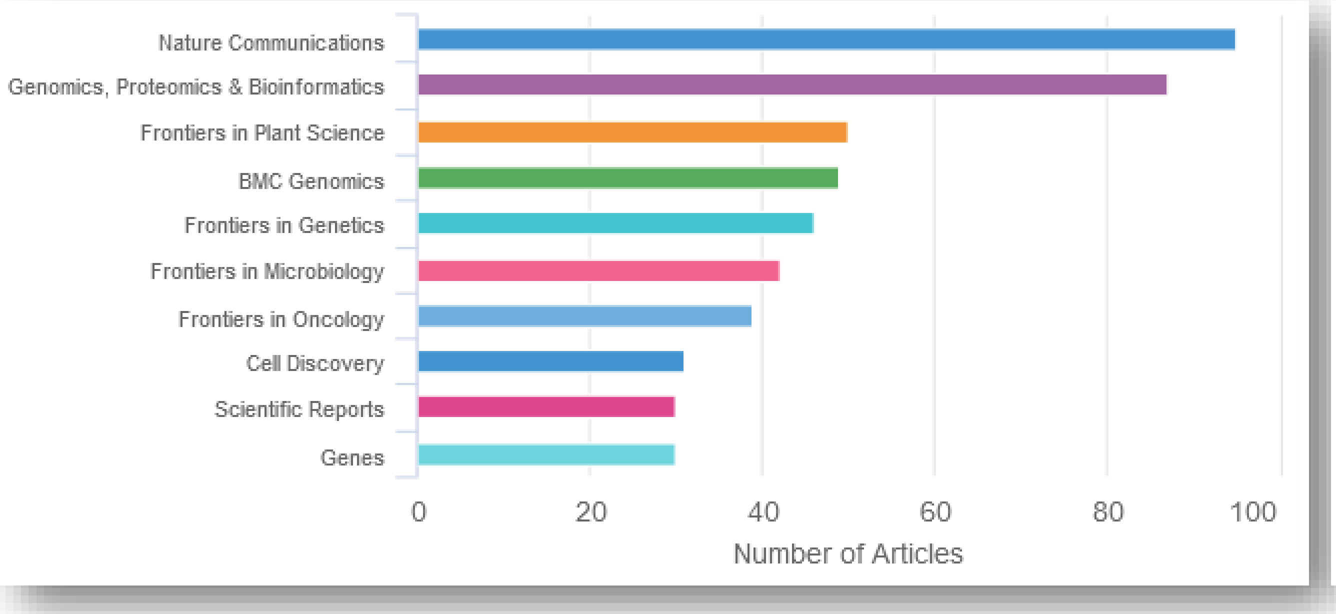
Search:

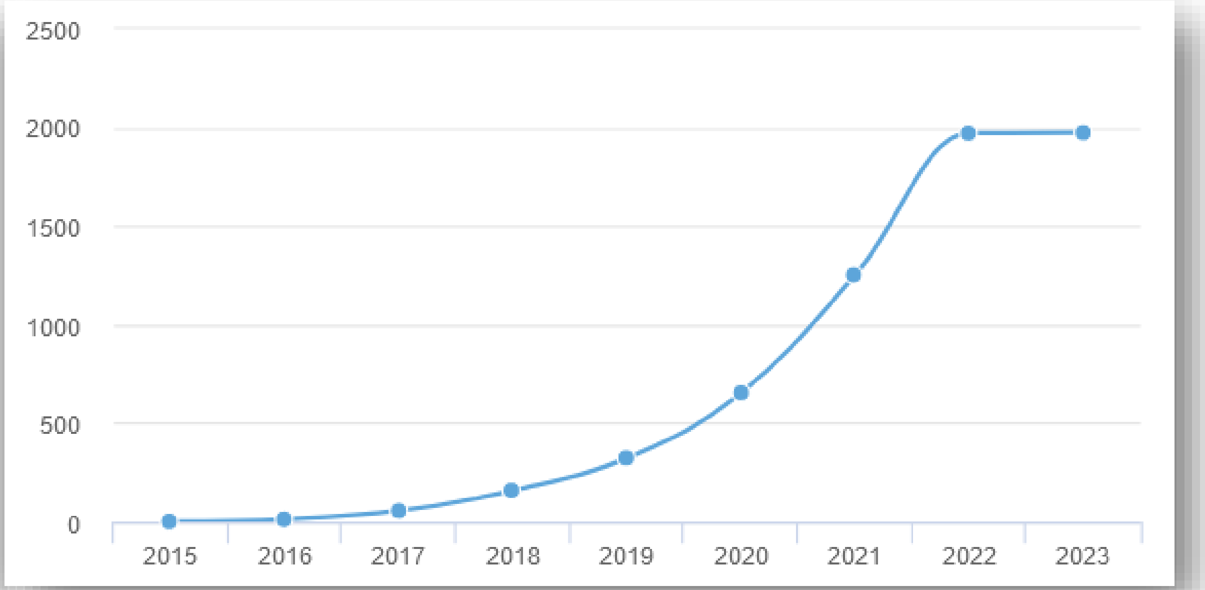
Agencies	Grants	GSAs	Experiments	Runs
Ministry of Science and Technology of the People's Republic of China (MOST)	729	507	27685	30577
National Natural Science Foundation of China (NSFC)	4552	1773	86747	93631
Chinese Academy of Sciences (CAS)	401	424	24197	27787
Others	5022	467	37995	38729

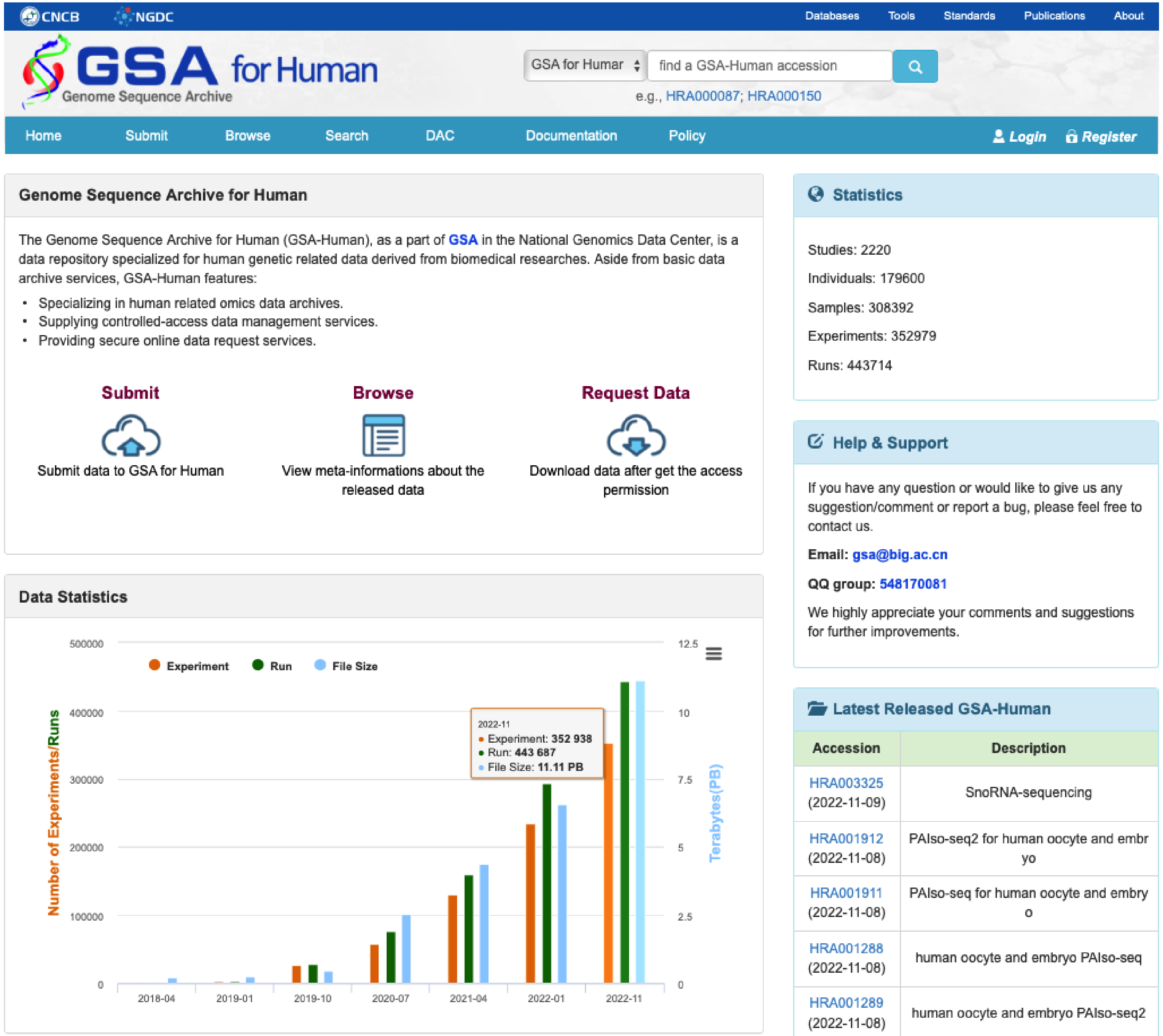
GSA Supported Publications

Total Journals: 457

Total Articles: ~2000







Genome Sequence Archive for Human

The Genome Sequence Archive for Human (GSA-Human), as a part of **GSA** in the National Genomics Data Center, is a data repository specialized for human genetic related data derived from biomedical researches. Aside from basic data archive services, GSA-Human features:

- Specializing in human related omics data archives.
- Supplying controlled-access data management services.
- Providing secure online data request services.

Submit

Submit data to GSA for Human

Browse

View meta-informations about the released data

Request Data

Download data after get the access permission

Data Statistics

Statistics

Studies: 2220

Individuals: 179600

Samples: 308392

Experiments: 352979

Runs: 443714

Help & Support

If you have any question or would like to give us any suggestion/comment or report a bug, please feel free to contact us.

Email: gsa@big.ac.cn

QQ group: 548170081

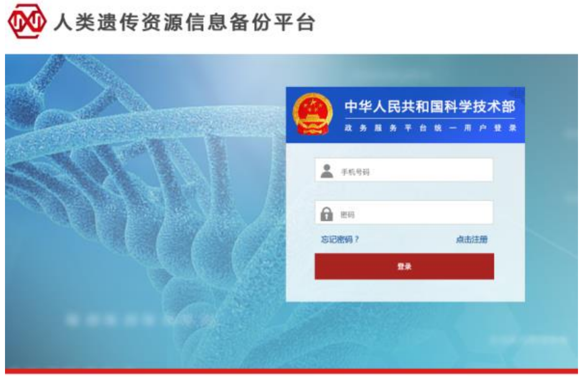
We highly appreciate your comments and suggestions for further improvements.

Latest Released GSA-Human

Accession	Description
HRA003325 (2022-11-09)	SnoRNA-sequencing
HRA001912 (2022-11-08)	PAIso-seq2 for human oocyte and embryo
HRA001911 (2022-11-08)	PAIso-seq for human oocyte and embryo
HRA001288 (2022-11-08)	human oocyte and embryo PAIso-seq
HRA001289 (2022-11-08)	human oocyte and embryo PAIso-seq2

Human Data Backup & Registration Protocol

Backup Center



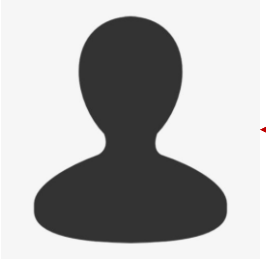
人类遗传资源信息备份平台

中华人民共和国科学技术部

政务服务平台统一用户登录

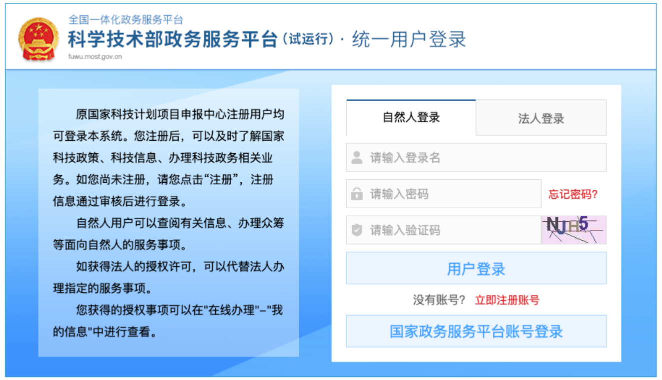
Backup

Backup ID



Researchers

Registration ID



全国一体化政务服务平台

科学技术部政务服务平台（试运行）· 统一用户登录

fuwu.most.gov.cn

原国家科技计划项目申报中心注册用户均可登录本系统。您注册后，可以及时了解国家科技政策、科技信息、办理科技政务相关业务。如您尚未注册，请您点击“注册”，注册信息通过审核后进行登录。

自然人用户可以查阅有关信息、办理众筹等面向自然人的服务事项。

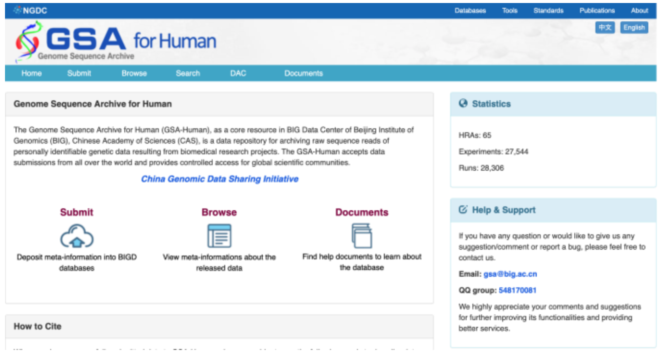
如获得法人的授权许可，可以代替法人办理指定的服务事项。

您获得的授权事项可以在"在线办理"–"我的信息"中进行查看。

Registration Center

Submission

Accession #s



Genome Sequence Archive for Human

The Genome Sequence Archive for Human (GSA-Human), as a core resource in BIG Data Center of Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS), is a data repository for archiving raw sequence reads of personally identifiable genetic data resulting from biomedical research projects. The GSA-Human accepts data submissions from all over the world and provides controlled access for global scientific communities.

China Genomic Data Sharing Initiative

Submit

Deposit meta-information into BIGD databases

Browse

View meta-informations about the released data

Documents

Find help documents to learn about the database

How to Cite

Statistics

HRAs: 65

Experiments: 27,544

Runs: 28,306

Help & Support

If you have any question or would like to give us any suggestion/comment or report a bug, please feel free to contact us.

Email: gsa@big.ac.cn

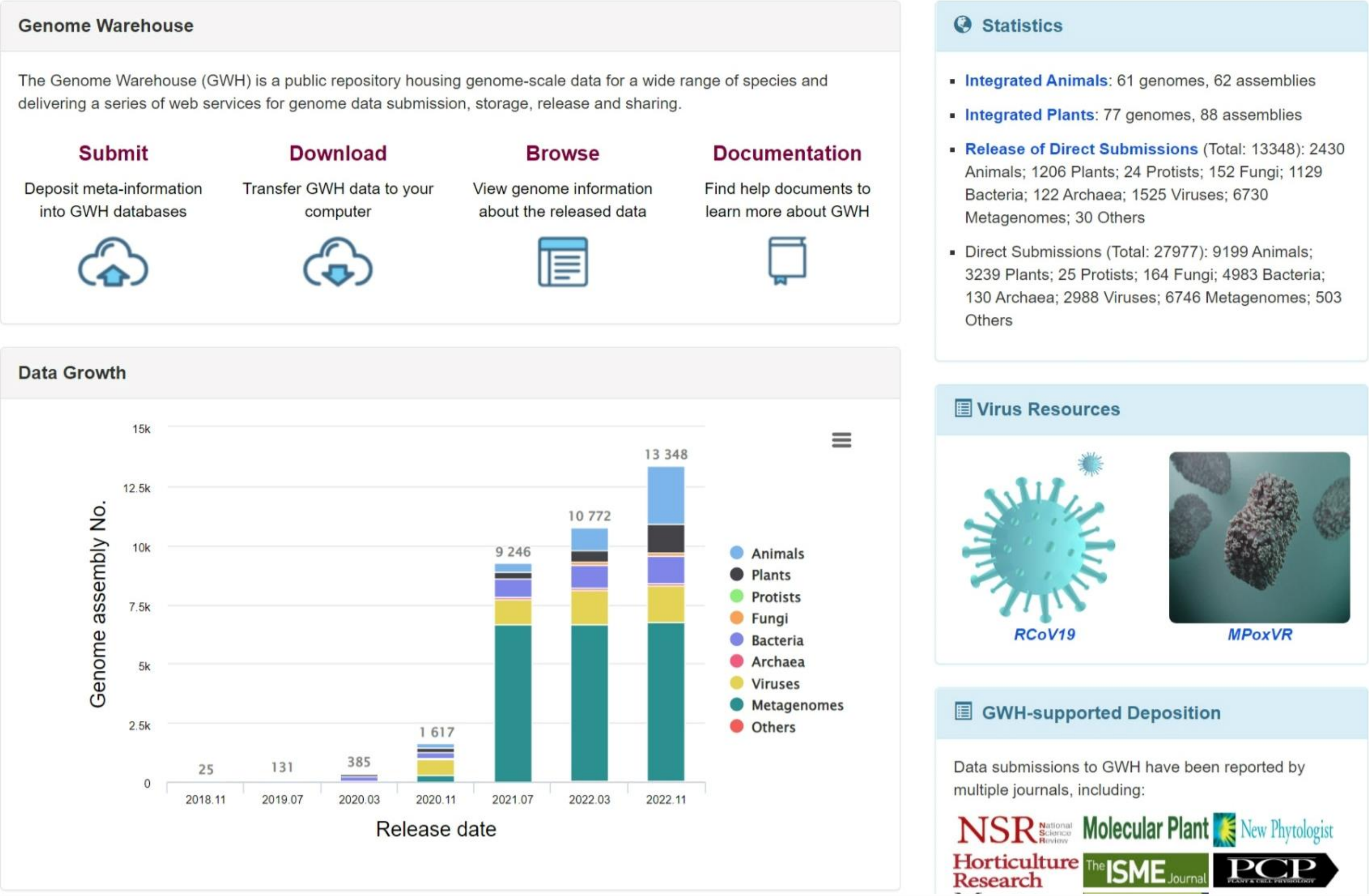
QQ group: 548170081

We highly appreciate your comments and suggestions for further improving its functionalities and providing better services.

CNCB-NGDC GSA-Human

Backup ID

Genome Warehouse - GWH



Genome Warehouse

The Genome Warehouse (GWH) is a public repository housing genome-scale data for a wide range of species and delivering a series of web services for genome data submission, storage, release and sharing.

Submit
Deposit meta-information into GWH databases

Download
Transfer GWH data to your computer

Browse
View genome information about the released data

Documentation
Find help documents to learn more about GWH

Statistics

- **Integrated Animals**: 61 genomes, 62 assemblies
- **Integrated Plants**: 77 genomes, 88 assemblies
- **Release of Direct Submissions** (Total: 13348): 2430 Animals; 1206 Plants; 24 Protists; 152 Fungi; 1129 Bacteria; 122 Archaea; 1525 Viruses; 6730 Metagenomes; 30 Others
- Direct Submissions (Total: 27977): 9199 Animals; 3239 Plants; 25 Protists; 164 Fungi; 4983 Bacteria; 130 Archaea; 2988 Viruses; 6746 Metagenomes; 503 Others

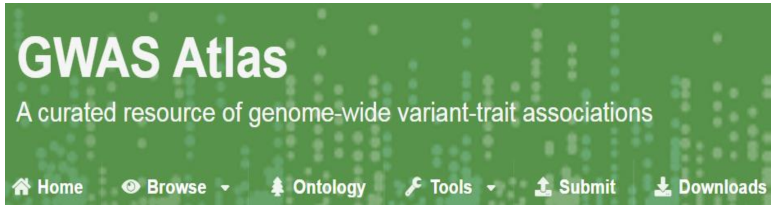
Data Growth

Virus Resources

GWH-supported Deposition

Data submissions to GWH have been reported by multiple journals, including:

Multi-omics association knowledgebases

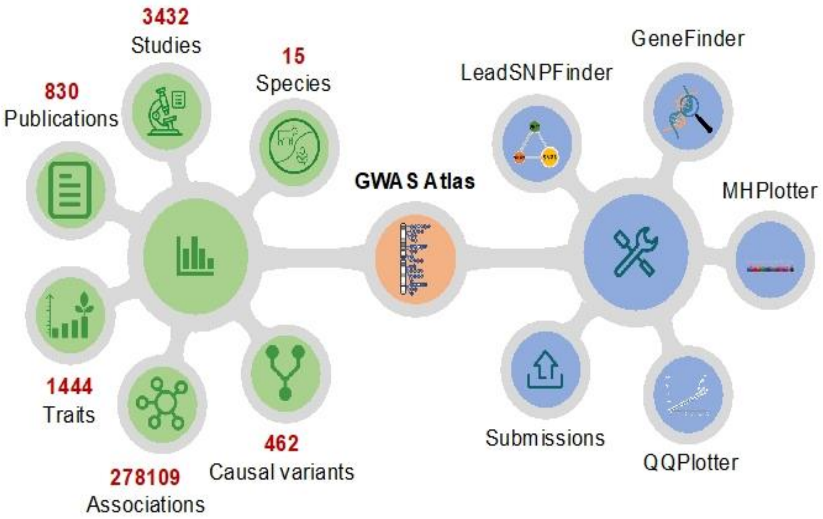

GWAS Atlas

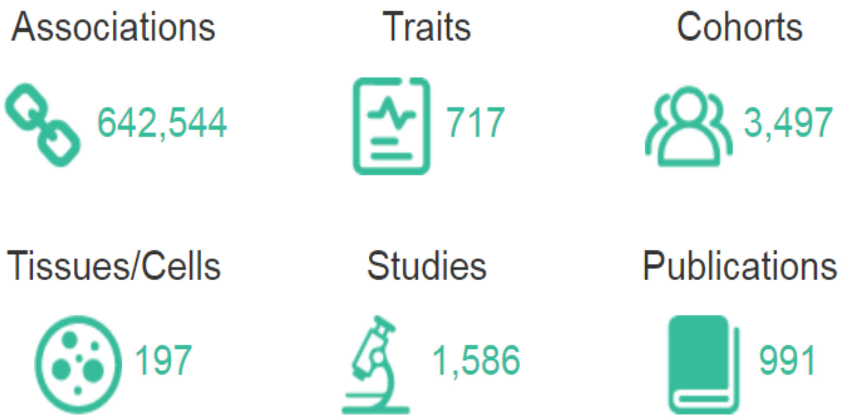
A curated resource of genome-wide variant-trait associations

EWAS Atlas @ EWAS Open Platform

A knowledgebase of epigenome-wide association studies

TWAS Atlas




Associations 642,544

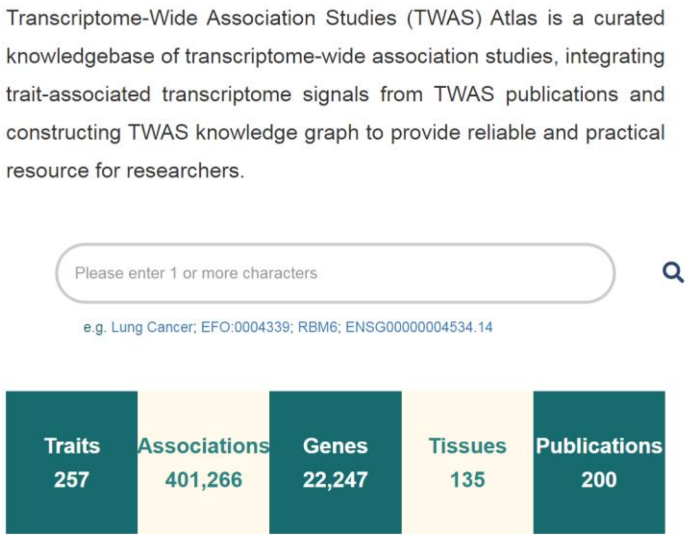
Traits 717

Cohorts 3,497

Tissues/Cells 197

Studies 1,586

Publications 991


Transcriptome-Wide Association Studies (TWAS) Atlas is a curated knowledgebase of transcriptome-wide association studies, integrating trait-associated transcriptome signals from TWAS publications and constructing TWAS knowledge graph to provide reliable and practical resource for researchers.

Please enter 1 or more characters
e.g. Lung Cancer; EFO:0004339; RBM6; ENSG00000004534.14

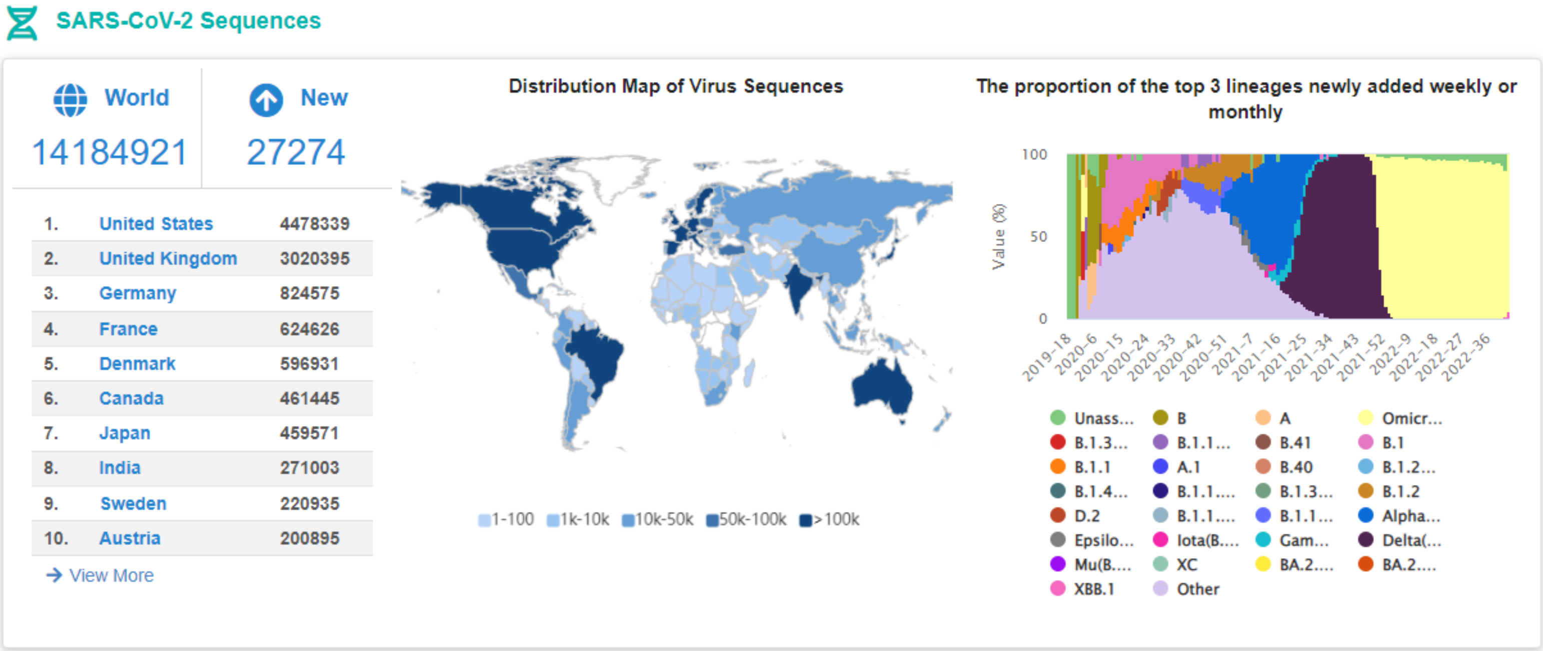
Traits	Associations	Genes	Tissues	Publications
257	401,266	22,247	135	200

Nucleic Acids Research, 2020, 2022

Nucleic Acids Research, 2019

Nucleic Acids Research, 2023

RCoV19



SARS-CoV-2 Sequences

World 14184921

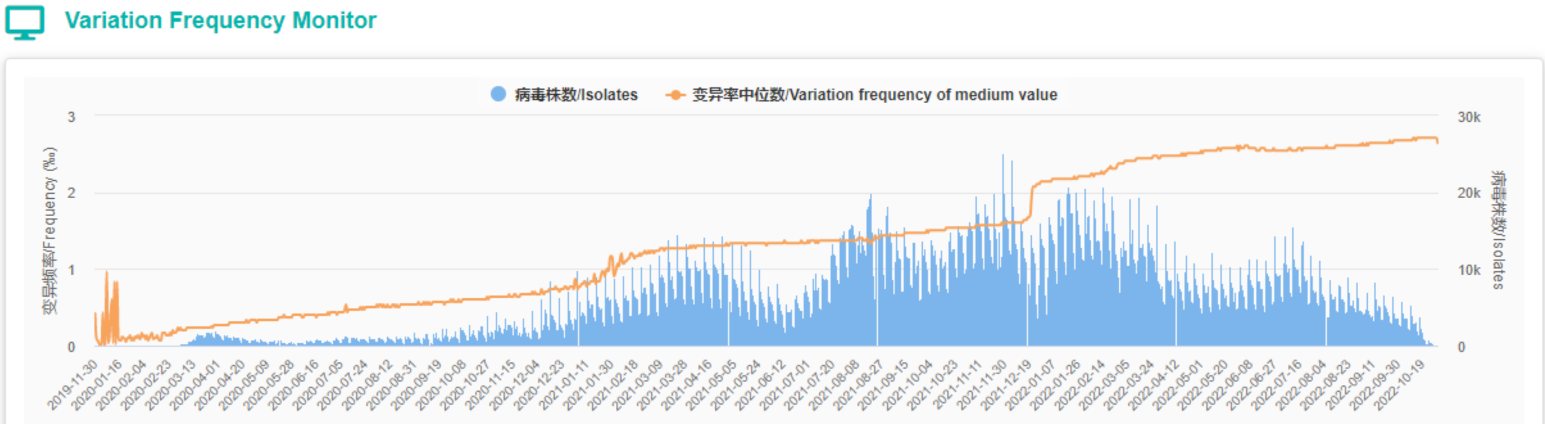
New 27274

1.	United States	4478339
2.	United Kingdom	3020395
3.	Germany	824575
4.	France	624626
5.	Denmark	596931
6.	Canada	461445
7.	Japan	459571
8.	India	271003
9.	Sweden	220935
10.	Austria	200895

→ View More

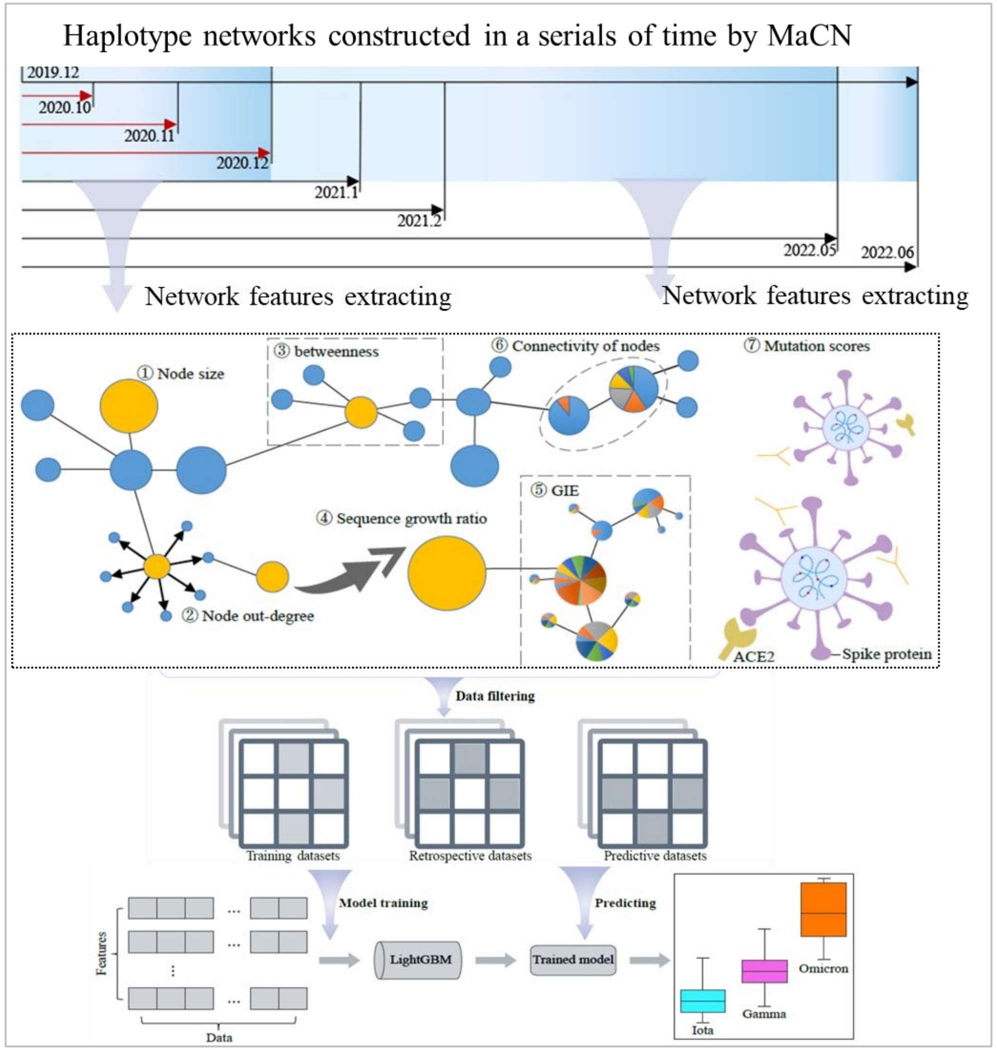
Distribution Map of Virus Sequences

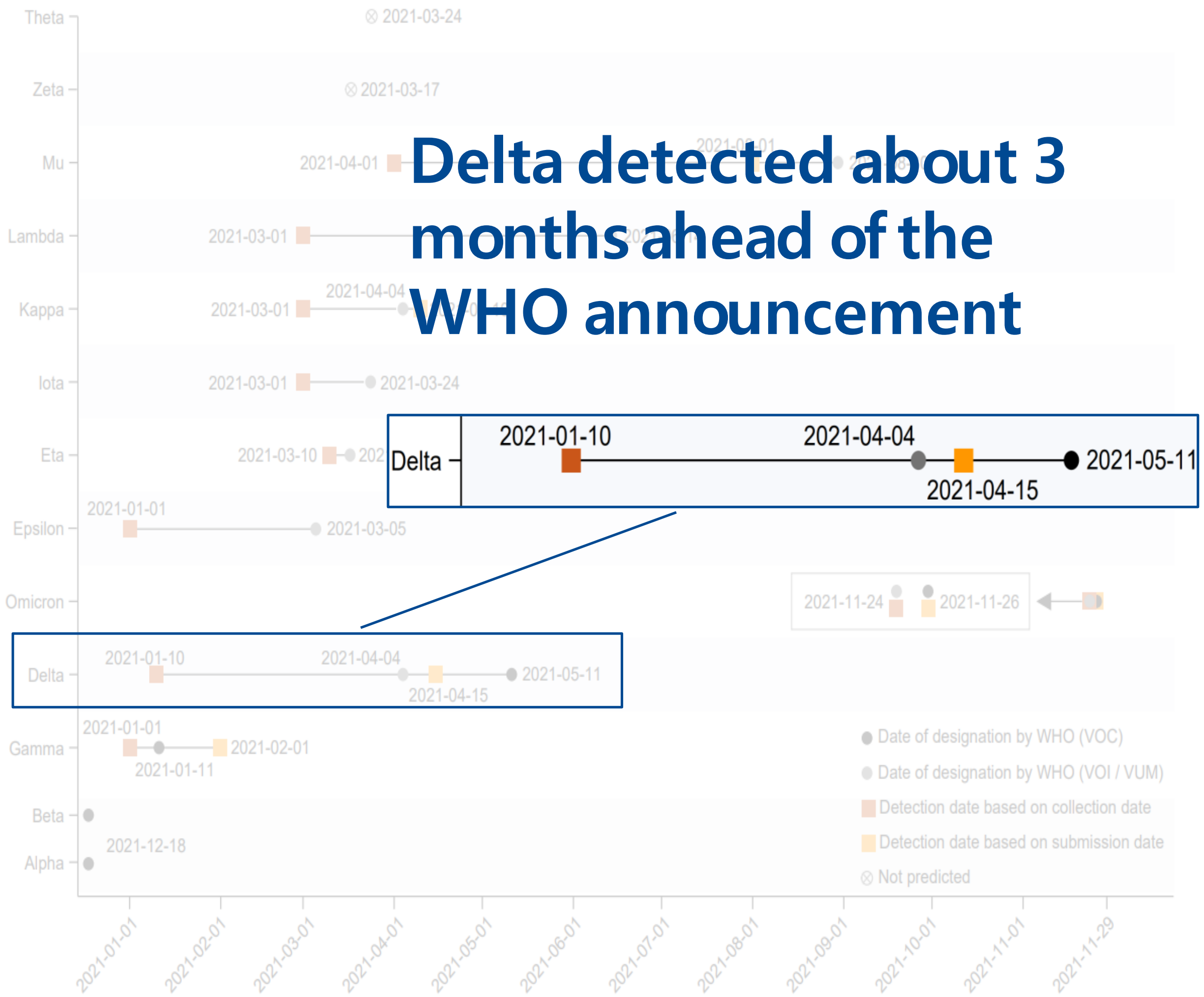
The proportion of the top 3 lineages newly added weekly or monthly



Variation Frequency Monitor

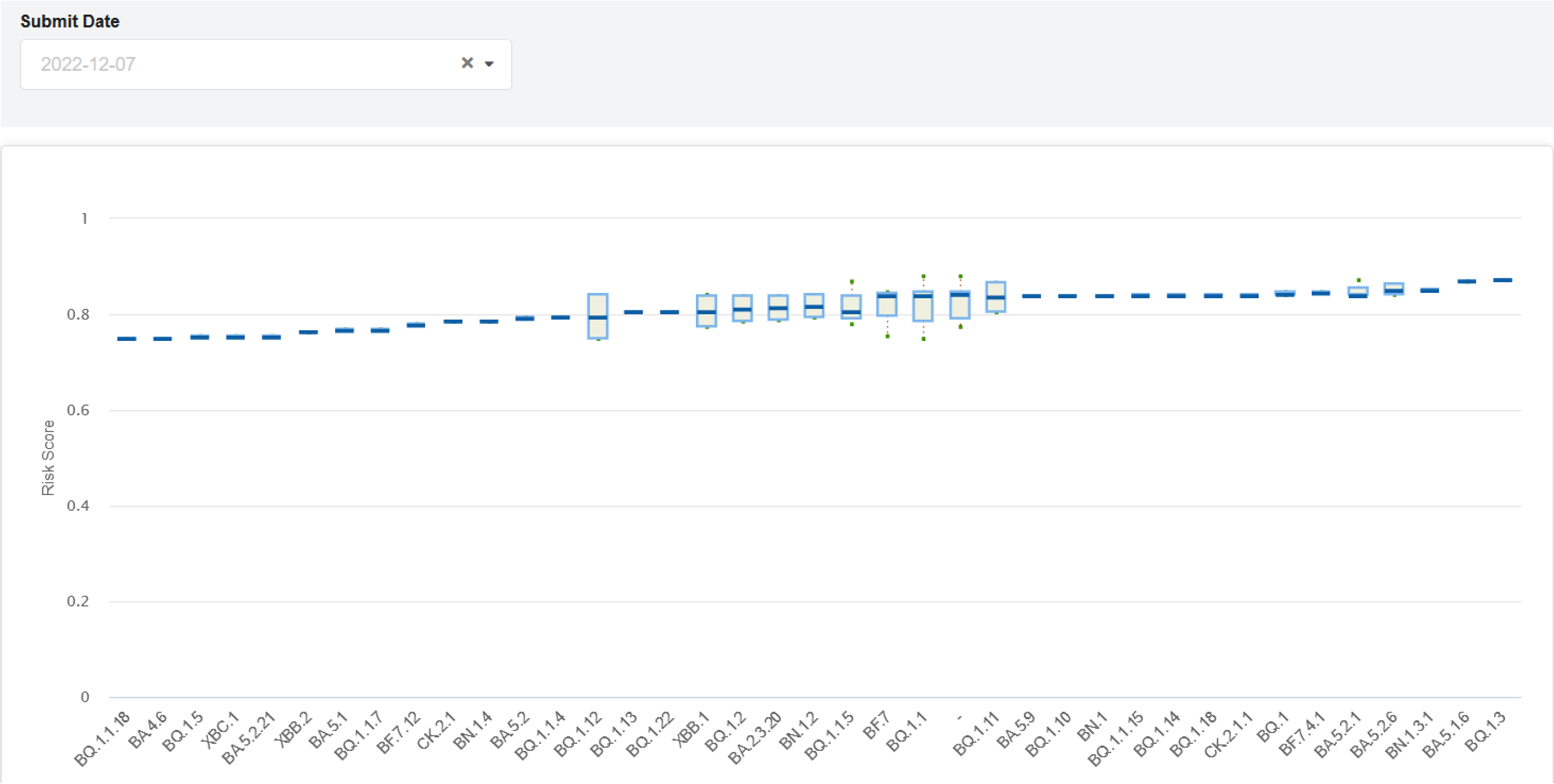
Machine learning detection of high-risk SARS-CoV-2 variants





Delta detected about 3 months ahead of the WHO announcement

Early-warning of high-risk variants

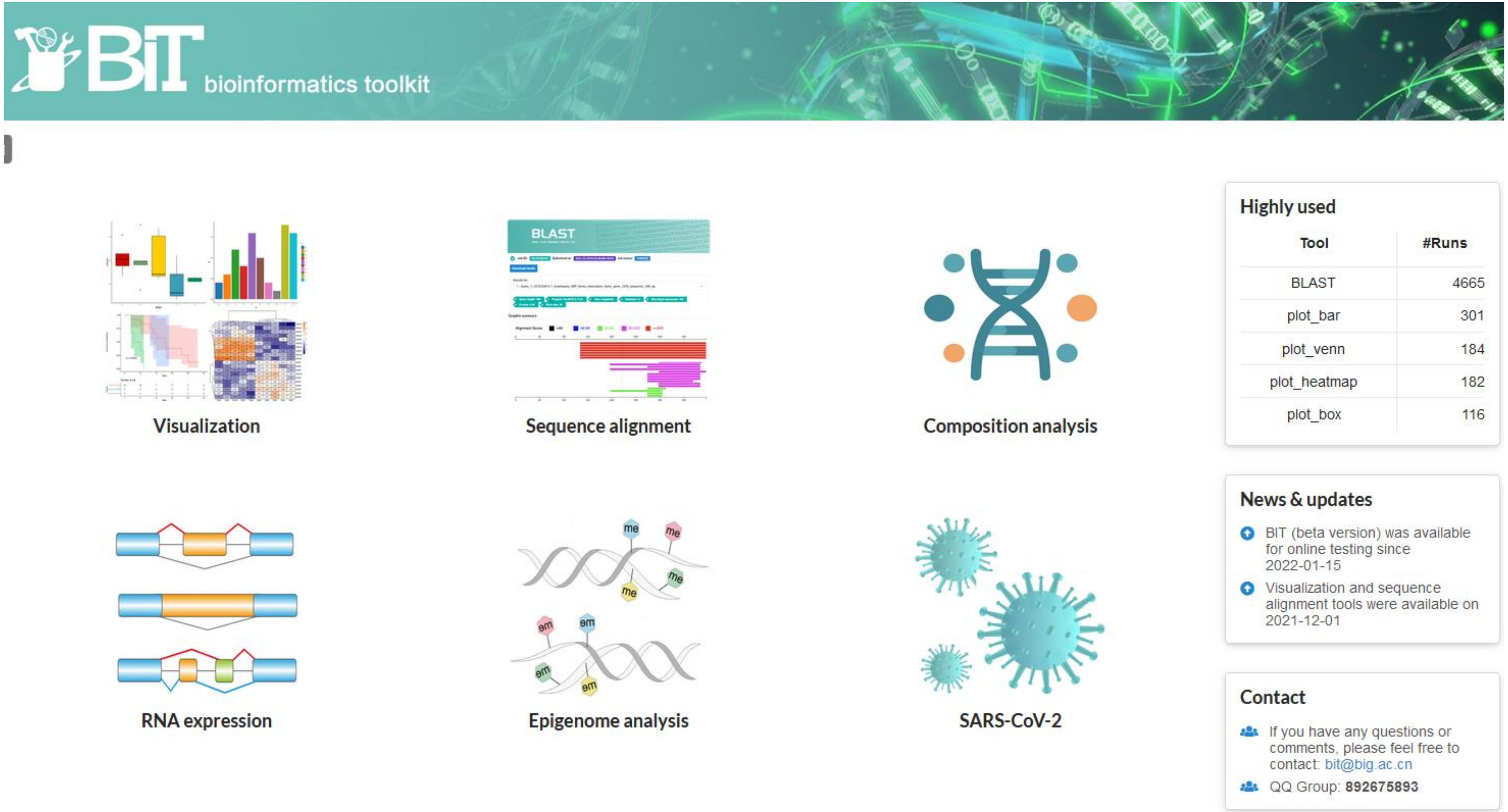


Submit Date

2022-12-07

https://ngdc.cncb.ac.cn/ncov/monitoring/risk

Bioinformatics toolkit - BIT



Highly used

Tool	#Runs
BLAST	4665
plot_bar	301
plot_venn	184
plot_heatmap	182
plot_box	116

News & updates

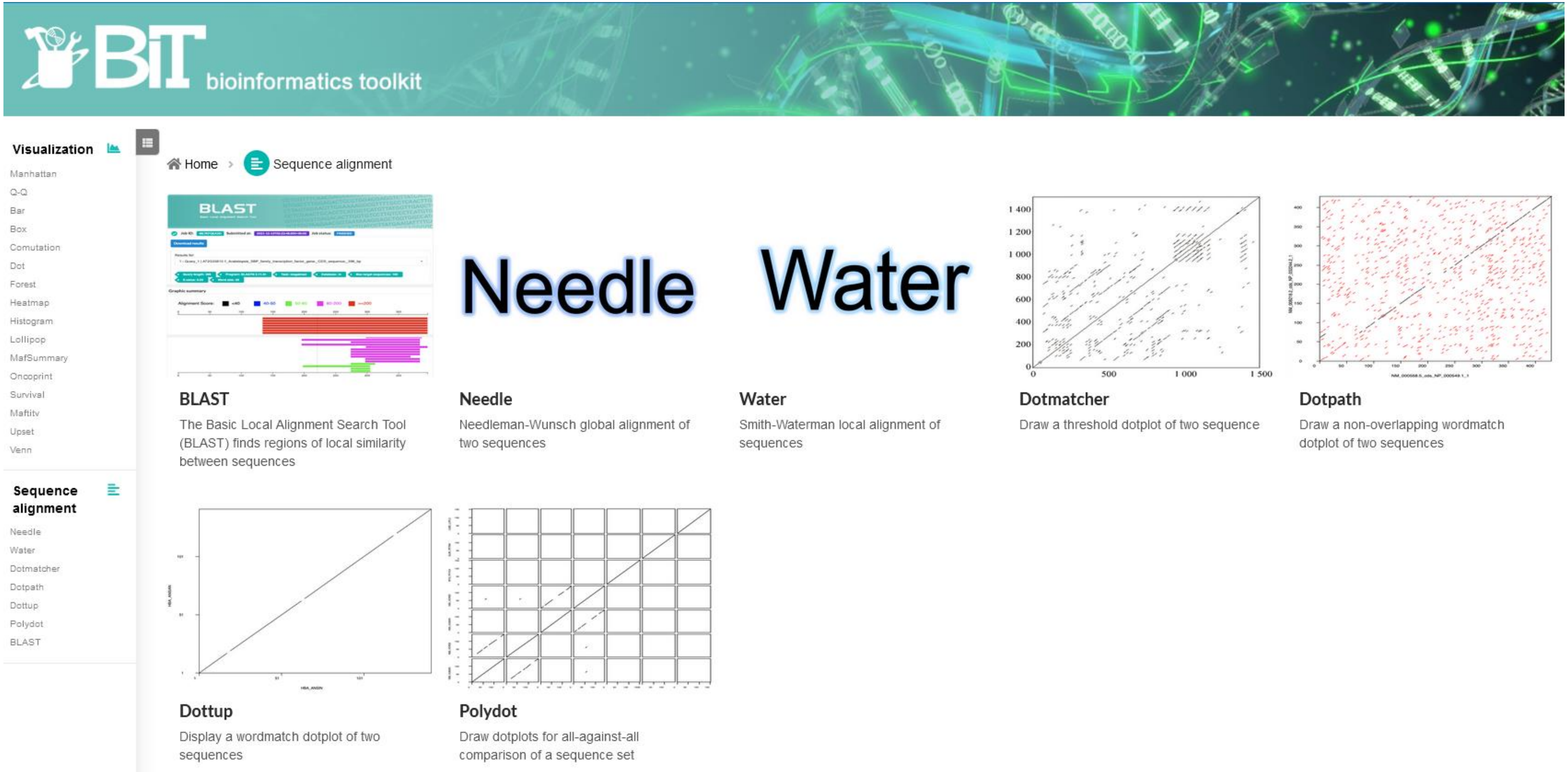
- BIT (beta version) was available for online testing since 2022-01-15
- Visualization and sequence alignment tools were available on 2021-12-01

Contact

- If you have any questions or comments, please feel free to contact: bit@big.ac.cn
- QQ Group: **892675893**

https://ngdc.cncb.ac.cn/bit/

Sequence Alignment Tools



BIT bioinformatics toolkit

Visualization

- Manhattan
- Q-Q
- Bar
- Box
- Comutation
- Dot
- Forest
- Heatmap
- Histogram
- Lollipop
- MafSummary
- Oncoprint
- Survival
- Maftitv
- Upset
- Venn

Sequence alignment

- Needle
- Water
- Dotmatcher
- Dotpath
- Dottup
- Polydot
- BLAST

Home > Sequence alignment

BLAST

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences

Needle

Needleman-Wunsch global alignment of two sequences

Water

Smith-Waterman local alignment of sequences

Dotmatcher

Draw a threshold dotplot of two sequence

Dotpath

Draw a non-overlapping wordmatch dotplot of two sequences

Dottup

Display a wordmatch dotplot of two sequences

Polydot

Draw dotplots for all-against-all comparison of a sequence set

https://ngdc.cncb.ac.cn/bit/seqaln

IT Infrastructure

Public Data Storage

- **24.5PB Storage capacity**
- **9.6PB Newly added storage**

Application Clusters

- **67 Application servers**
- **11 Newly added servers**

Backup

- **15.6PB Archive backup capacity**
- **Tape library system 2 sets**
- **2.4PB Newly added capacity**

Internet Bandwidth

- **4Gb Data center Internet bandwidth**
- **The average bandwidth usage is more than 70%**

HPC

- **210 Computing Nodes**
- **8900 CPU computing cores**
- **269 TFlops CPUs' computational ability**
- **82 TFlops GPUs' computational ability**
- **8 Newly added 512GB memory Nodes**

HP Storage

- **8.2PB Storage capacity**

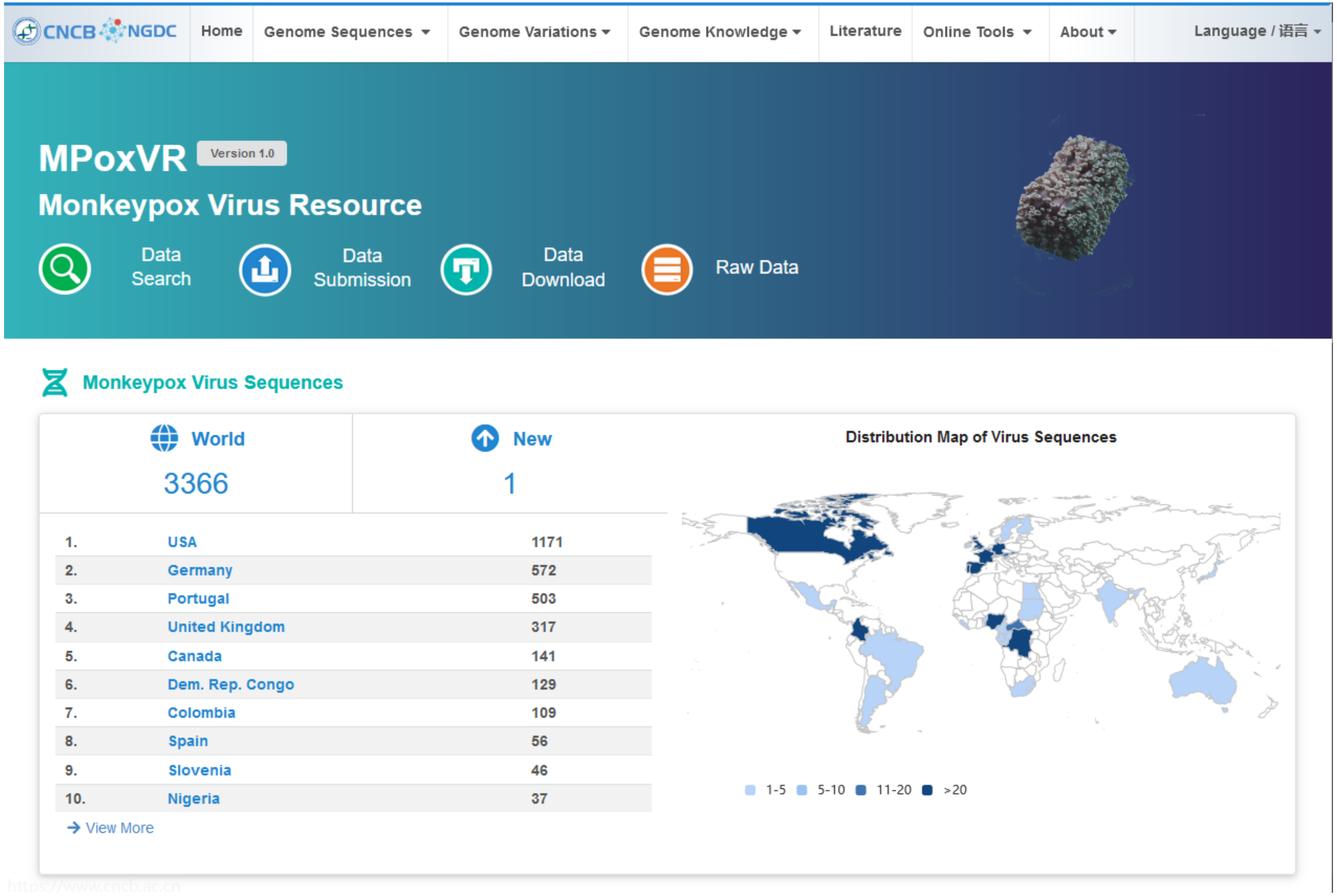
CONTENTS

1 Updates

2 Development

3 Future directions

Monkeypox Virus Resource



MPoxVR Version 1.0

Monkeypox Virus Resource

Monkeypox Virus Sequences

World 3366

New 1

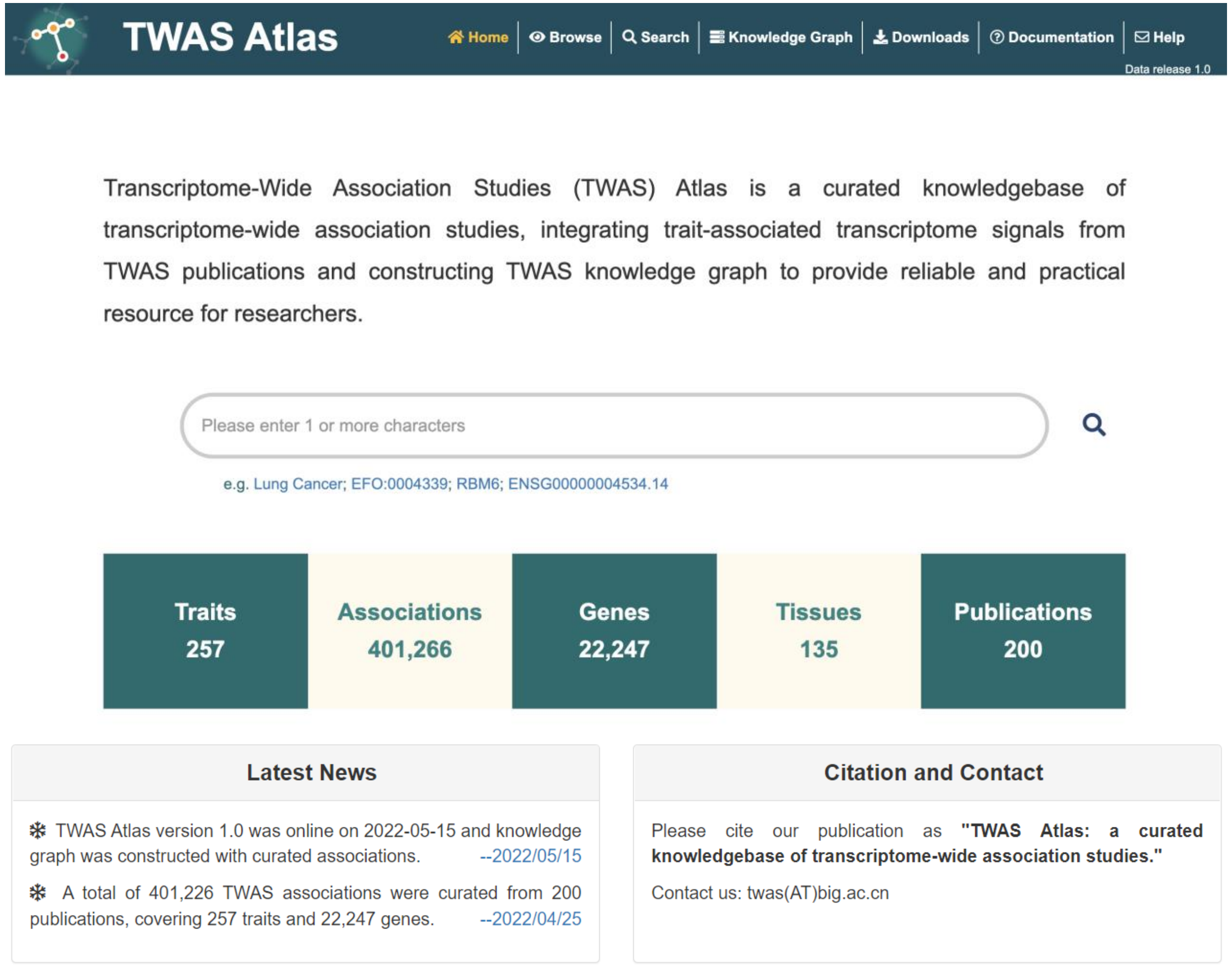
1.	USA	1171
2.	Germany	572
3.	Portugal	503
4.	United Kingdom	317
5.	Canada	141
6.	Dem. Rep. Congo	129
7.	Colombia	109
8.	Spain	56
9.	Slovenia	46
10.	Nigeria	37

Distribution Map of Virus Sequences

https://www.cncb.ac.cn

TWAS Atlas

a curated knowledgebase of transcriptome-wide association studies



TWAS Atlas

Home | Browse | Search | Knowledge Graph | Downloads | Documentation | Help

Data release 1.0

Transcriptome-Wide Association Studies (TWAS) Atlas is a curated knowledgebase of transcriptome-wide association studies, integrating trait-associated transcriptome signals from TWAS publications and constructing TWAS knowledge graph to provide reliable and practical resource for researchers.

Please enter 1 or more characters

e.g. Lung Cancer; EFO:0004339; RBM6; ENSG00000004534.14

Traits	Associations	Genes	Tissues	Publications
257	401,266	22,247	135	200

Latest News

❇ TWAS Atlas version 1.0 was online on 2022-05-15 and knowledge graph was constructed with curated associations. --2022/05/15

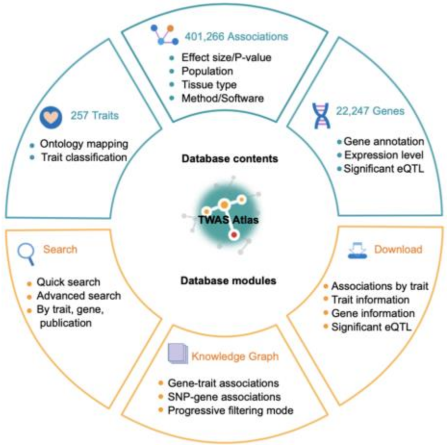
❇ A total of 401,226 TWAS associations were curated from 200 publications, covering 257 traits and 22,247 genes. --2022/04/25

Citation and Contact

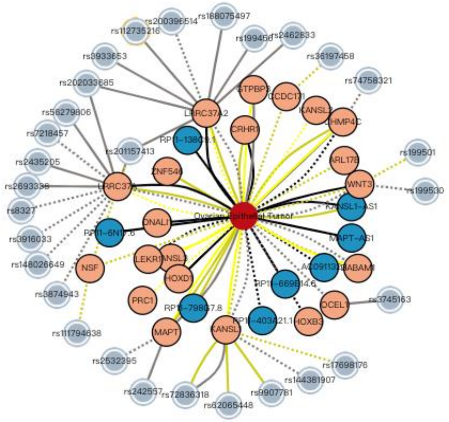
Please cite our publication as **"TWAS Atlas: a curated knowledgebase of transcriptome-wide association studies."**

Contact us: twas(AT)big.ac.cn

Overview of TWAS Atlas

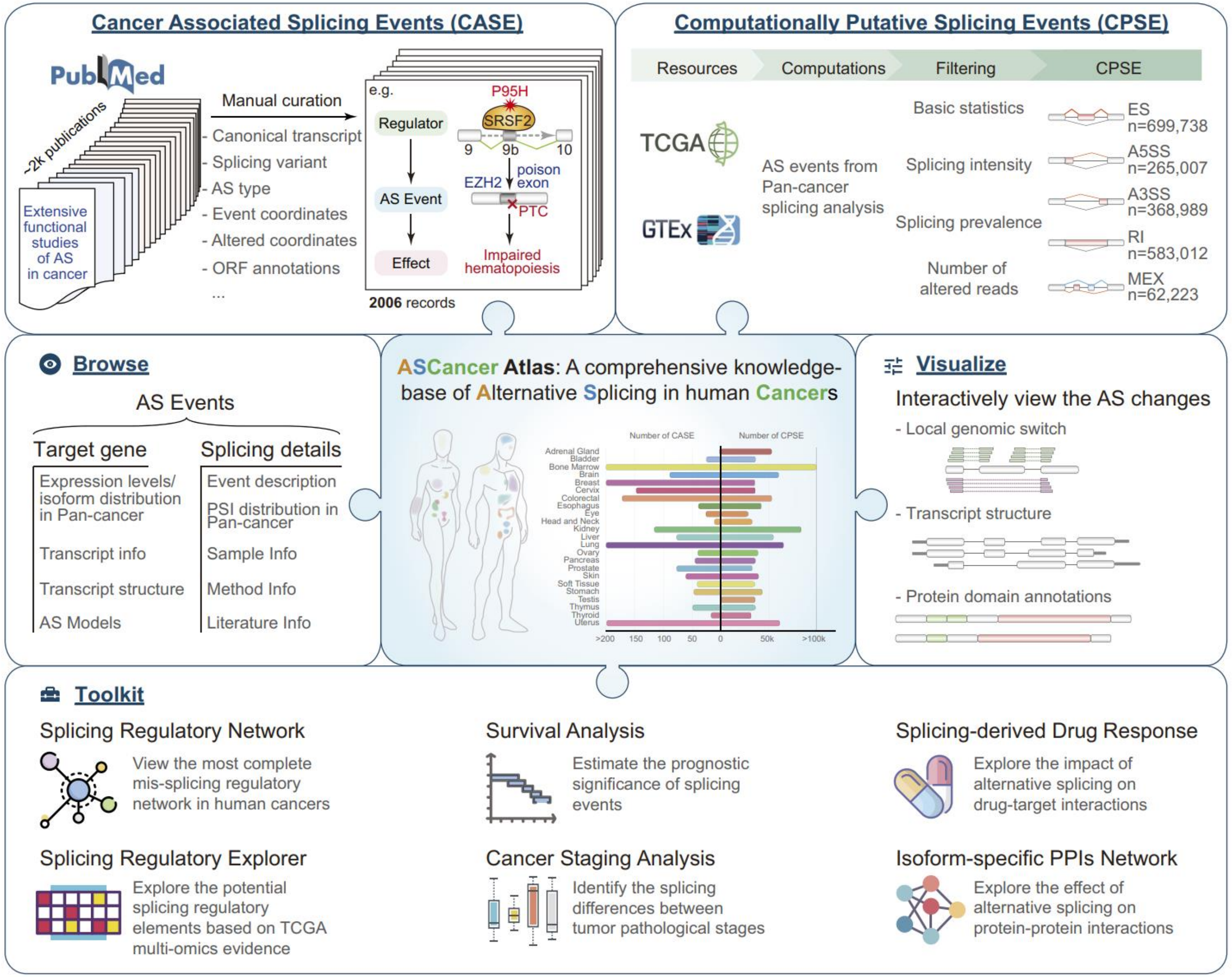


Ovarian Epithelial Tumor



***Nucleic Acids Research* (2023)**

ASCancer Atlas



Cancer Associated Splicing Events (CASE)

Computationally Putative Splicing Events (CPSE)

Browse

AS Events

Target gene	Splicing details
Expression levels/ isoform distribution in Pan-cancer	Event description
	PSI distribution in Pan-cancer
Transcript info	Sample Info
Transcript structure	Method Info
AS Models	Literature Info

ASCancer Atlas: A comprehensive knowledge-base of Alternative Splicing in human Cancers

Visualize

Interactively view the AS changes

- Local genomic switch
- Transcript structure
- Protein domain annotations

Toolkit

Splicing Regulatory Network

View the most complete mis-splicing regulatory network in human cancers

Survival Analysis

Estimate the prognostic significance of splicing events

Splicing-derived Drug Response

Explore the impact of alternative splicing on drug-target interactions

Splicing Regulatory Explorer

Explore the potential splicing regulatory elements based on TCGA multi-omics evidence

Cancer Staging Analysis

Identify the splicing differences between tumor pathological stages

Isoform-specific PPIs Network

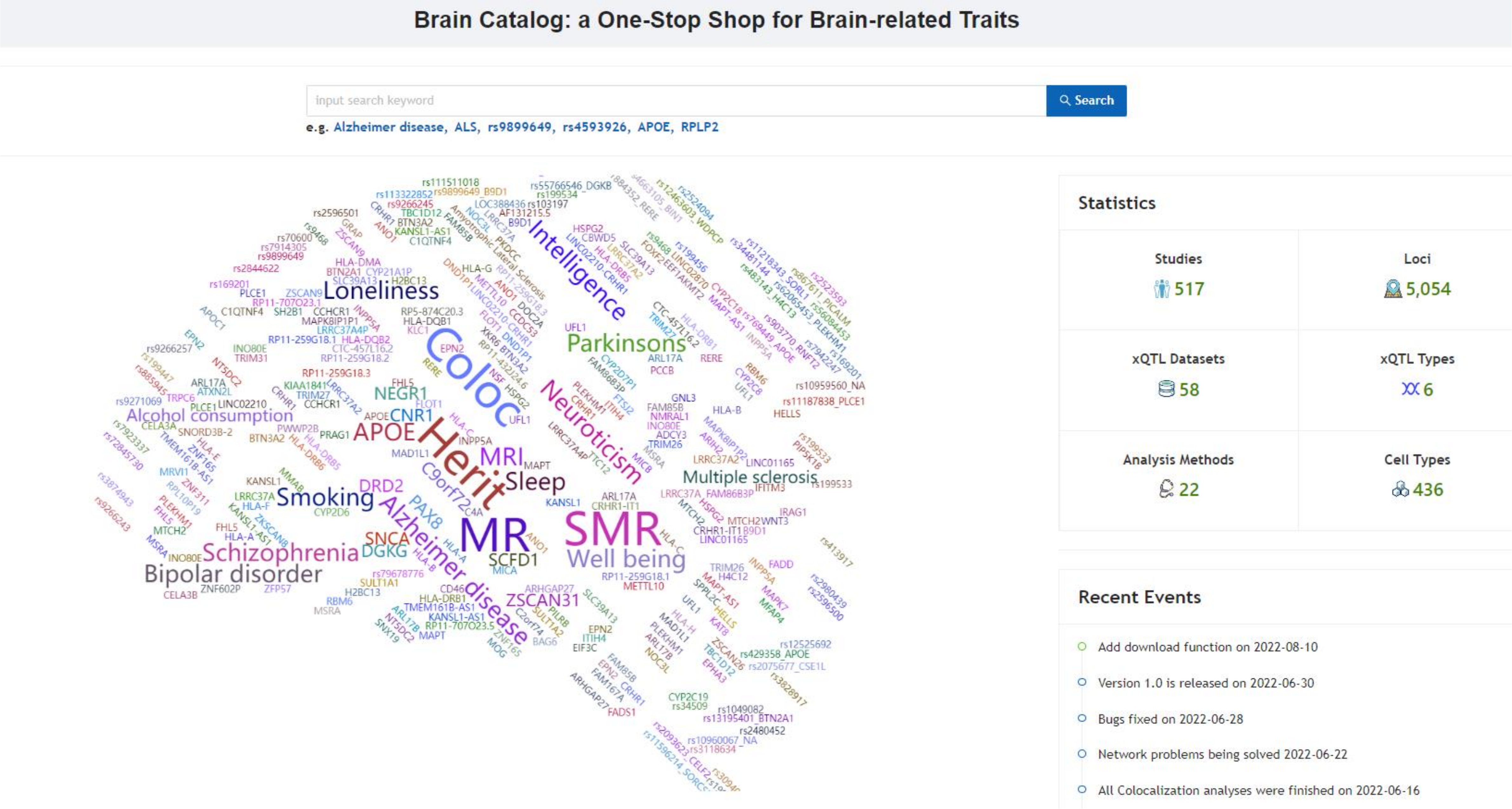
Explore the effect of alternative splicing on protein-protein interactions

Key features

- A collection of 2,006 experimentally validated cancer-associated splicing events(CASE)
- Each CASE records the complete information of upstream splicing regulators, splicing event annotations, downstream oncogenic effects and potential treatment targets
- The most complete splicing regulatory network so far
- An interactive splicing visualization tool and a suit of multi-dimensional online splicing analysis tools

***Nucleic Acids Research* (2023)**

https://ngdc.cncb.ac.cn/ascancer/



Brain Catalog: a One-Stop Shop for Brain-related Traits

e.g. Alzheimer disease, ALS, rs9899649, rs4593926, APOE, RPLP2

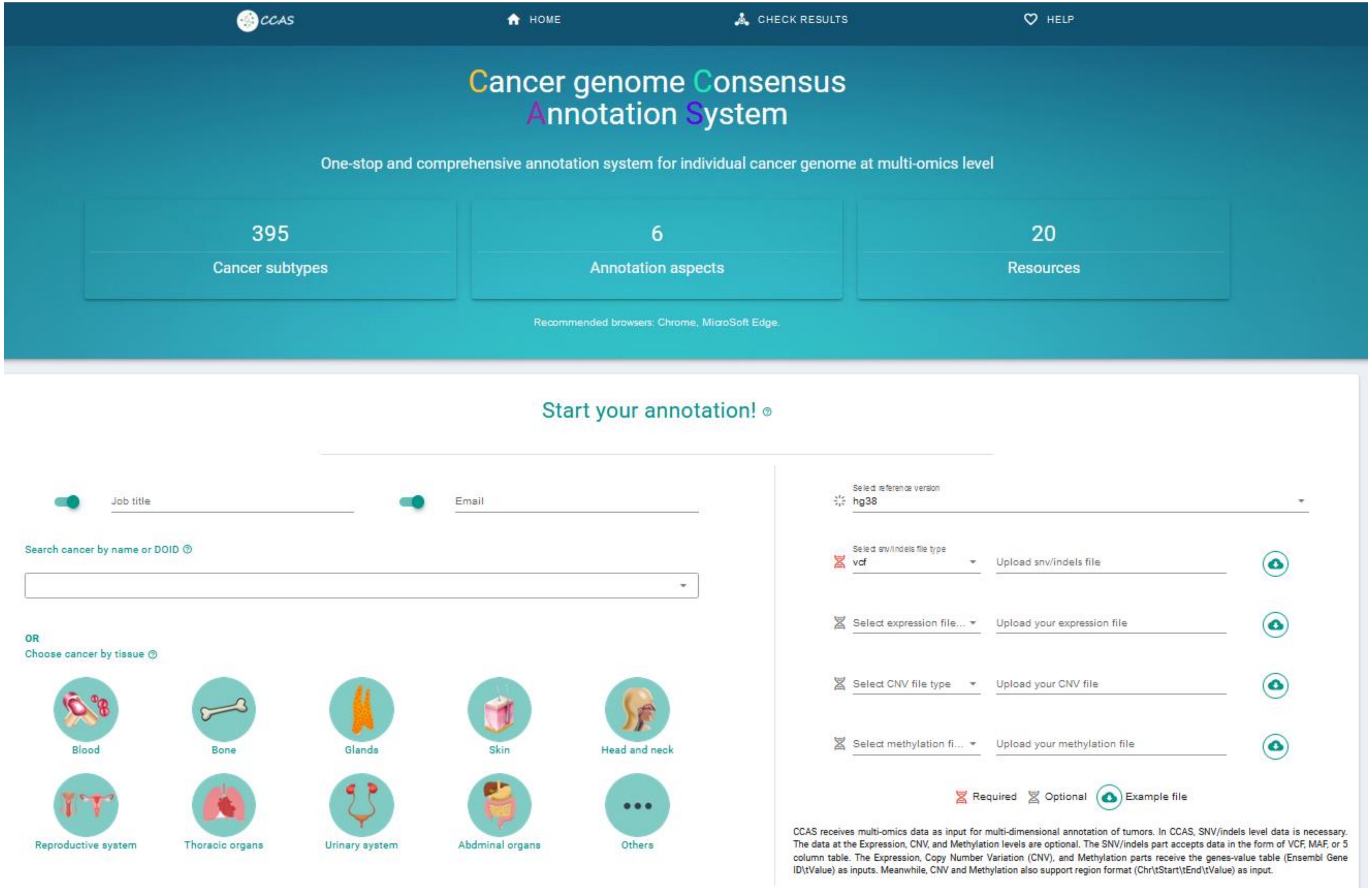
Statistics

Studies	Loci
517	5,054
xQTL Datasets	**xQTL Types**
58	6
Analysis Methods	**Cell Types**
22	436

Recent Events

- ○ Add download function on 2022-08-10
- ○ Version 1.0 is released on 2022-06-30
- ○ Bugs fixed on 2022-06-28
- ○ Network problems being solved 2022-06-22
- ○ All Colocalization analyses were finished on 2022-06-16

CCAS



Cancer genome Consensus Annotation System

One-stop and comprehensive annotation system for individual cancer genome at multi-omics level

395	6	20
Cancer subtypes	Annotation aspects	Resources

Recommended browsers: Chrome, MicroSoft Edge.

Start your annotation!

Job title

Email

Search cancer by name or DOID

OR

Choose cancer by tissue

Select reference version: hg38

Select snv/indels file type: vcf — Upload snv/indels file

Select expression file... — Upload your expression file

Select CNV file type — Upload your CNV file

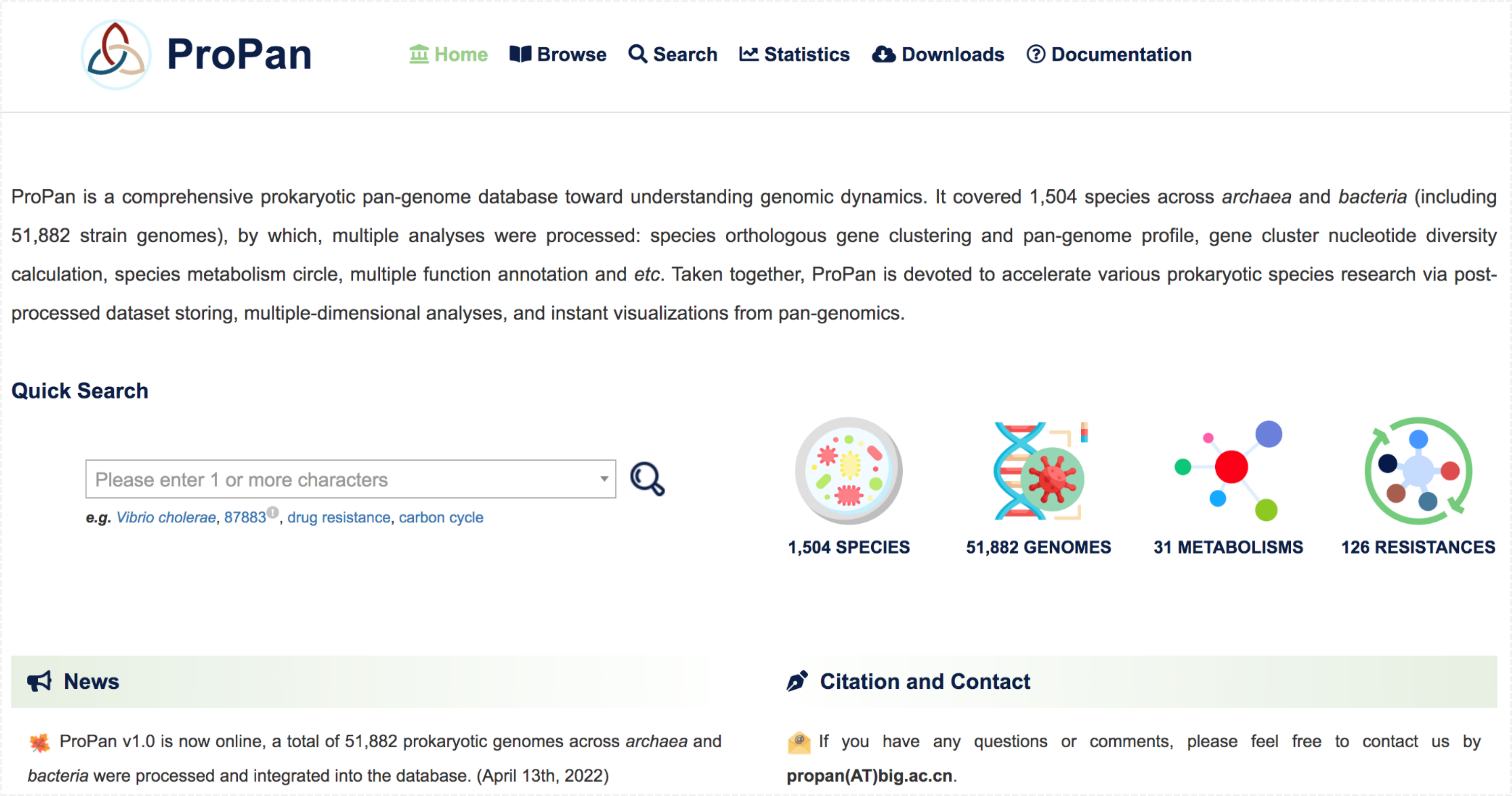
Select methylation fi... — Upload your methylation file

Required Optional Example file

CCAS receives multi-omics data as input for multi-dimensional annotation of tumors. In CCAS, SNV/indels level data is necessary. The data at the Expression, CNV, and Methylation levels are optional. The SNV/indels part accepts data in the form of VCF, MAF, or 5 column table. The Expression, Copy Number Variation (CNV), and Methylation parts receive the genes-value table (Ensembl Gene ID\tValue) as inputs. Meanwhile, CNV and Methylation also support region format (Chr\tStart\tEnd\tValue) as input.

ProPan

a comprehensive database for profiling prokaryotic pan-genome dynamics



ProPan is a comprehensive prokaryotic pan-genome database toward understanding genomic dynamics. It covered 1,504 species across *archaea* and *bacteria* (including 51,882 strain genomes), by which, multiple analyses were processed: species orthologous gene clustering and pan-genome profile, gene cluster nucleotide diversity calculation, species metabolism circle, multiple function annotation and *etc*. Taken together, ProPan is devoted to accelerate various prokaryotic species research via post-processed dataset storing, multiple-dimensional analyses, and instant visualizations from pan-genomics.

Quick Search

Please enter 1 or more characters

e.g. *Vibrio cholerae*, 87883, drug resistance, carbon cycle

News

ProPan v1.0 is now online, a total of 51,882 prokaryotic genomes across *archaea* and *bacteria* were processed and integrated into the database. (April 13th, 2022)

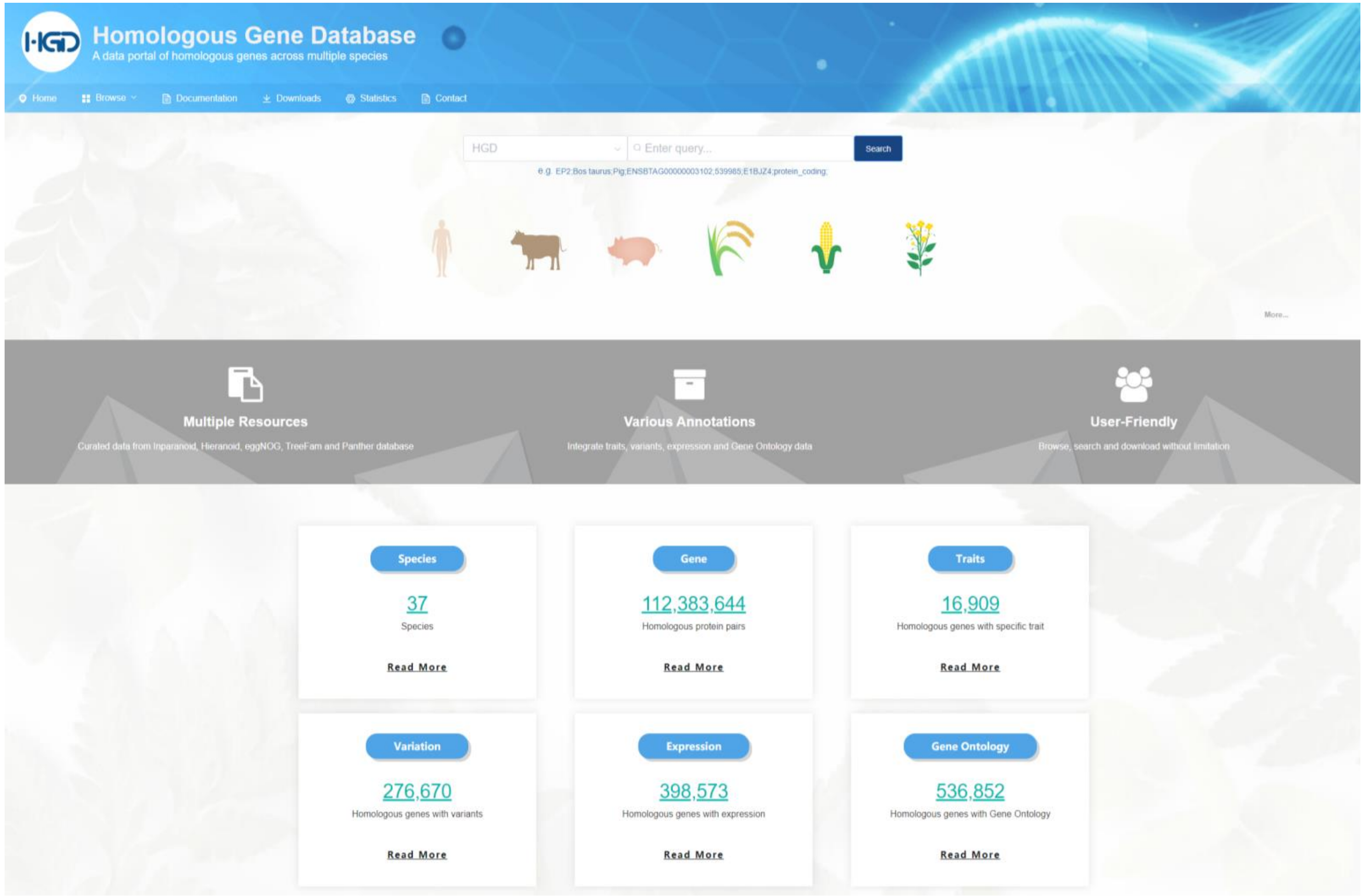
Citation and Contact

If you have any questions or comments, please feel free to contact us by **propan(AT)big.ac.cn**.

***Nucleic Acids Research* (2023)**

https://ngdc.cncb.ac.cn/propan

Homologous Gene Database - HGD



Homologous Gene Database

A data portal of homologous genes across multiple species

e.g. EP2;Bos taurus;Pig;ENSBTAG00000003102;539985;E1BJZ4;protein_coding;

Multiple Resources

Curated data from Inparanoid, Hieranoid, eggNOG, TreeFam and Panther database

Various Annotations

Integrate traits, variants, expression and Gene Ontology data

User-Friendly

Browse, search and download without limitation

Species	Gene	Traits
37 Species Read More	112,383,644 Homologous protein pairs Read More	16,909 Homologous genes with specific trait Read More

Variation	Expression	Gene Ontology
276,670 Homologous genes with variants Read More	398,573 Homologous genes with expression Read More	536,852 Homologous genes with Gene Ontology Read More

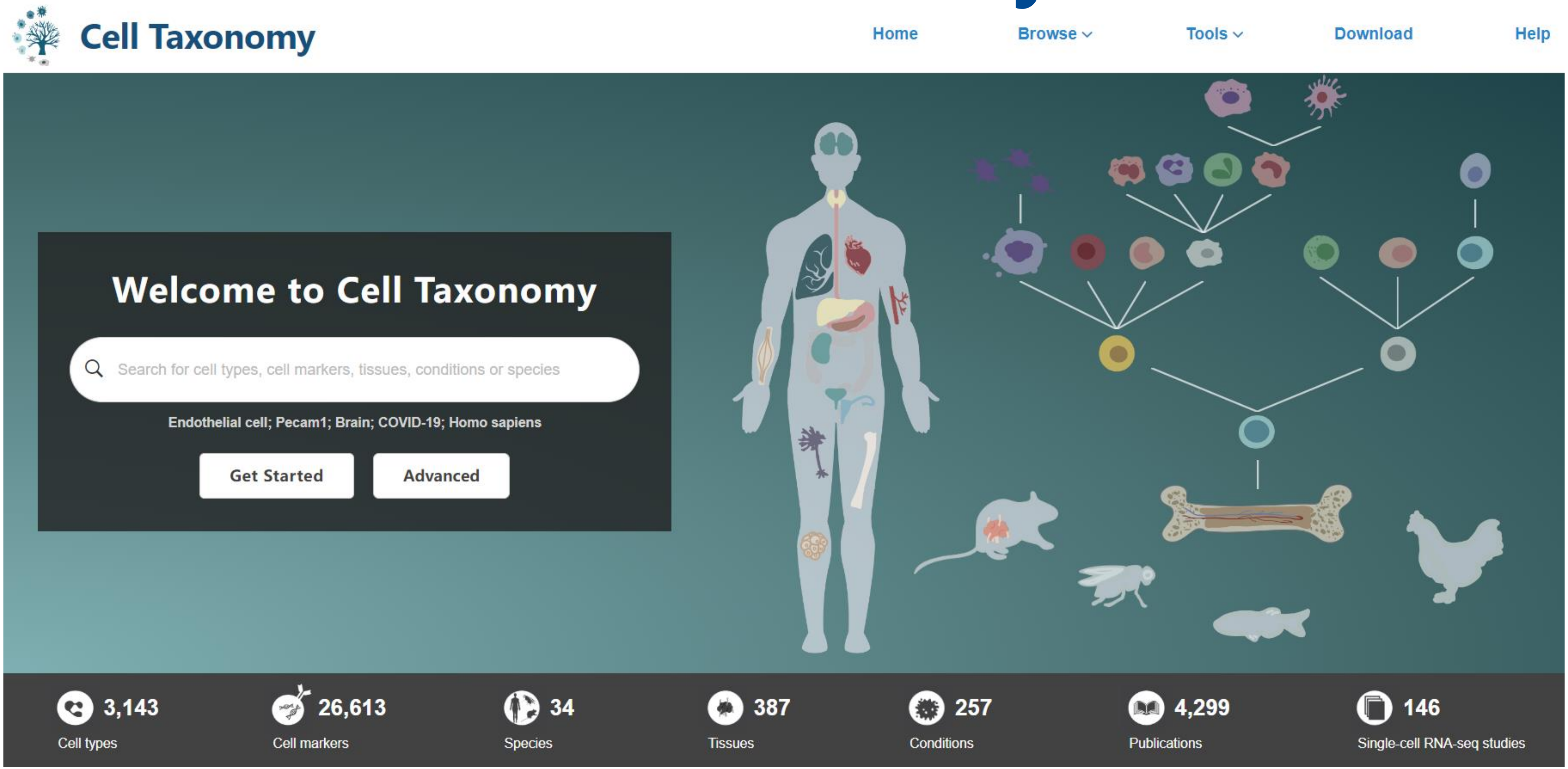
Nucleic Acids Research **(2023)**

Features:

- Integrated **112,383,644** homologous pairs from multiple resources
- Homologous genes with various annotations
  - **16,909** homologs with traits
  - **276,607** homologs with variants
  - **398,573** homologs with expression
  - **536,852** homologs with Gene Ontology
- Support a comparison function of homologs across multiple species

https://ngdc.cncb.ac.cn/hgd/

Cell Taxonomy



Cell Taxonomy

Welcome to Cell Taxonomy

Search for cell types, cell markers, tissues, conditions or species

Endothelial cell; Pecam1; Brain; COVID-19; Homo sapiens

3,143	26,613	34	387	257	4,299	146
Cell types	Cell markers	Species	Tissues	Conditions	Publications	Single-cell RNA-seq studies

Nucleic Acids Research **(2023)**

GSA’ s Application for the GCBR-Selection (the Global Core Biodata Resource Selection)

- Initiated by the Global Biodata Coalition, aiming to define Global Core Biodata Resources across biological, life science, and biomedical data resources (biodata resources) worldwide.
- GSA had passed the first round of the selection, and was invited to submit the full application.
- Waiting for the final result on 12th, December, 2022.

GLOBAL
BIODATA
COALITION

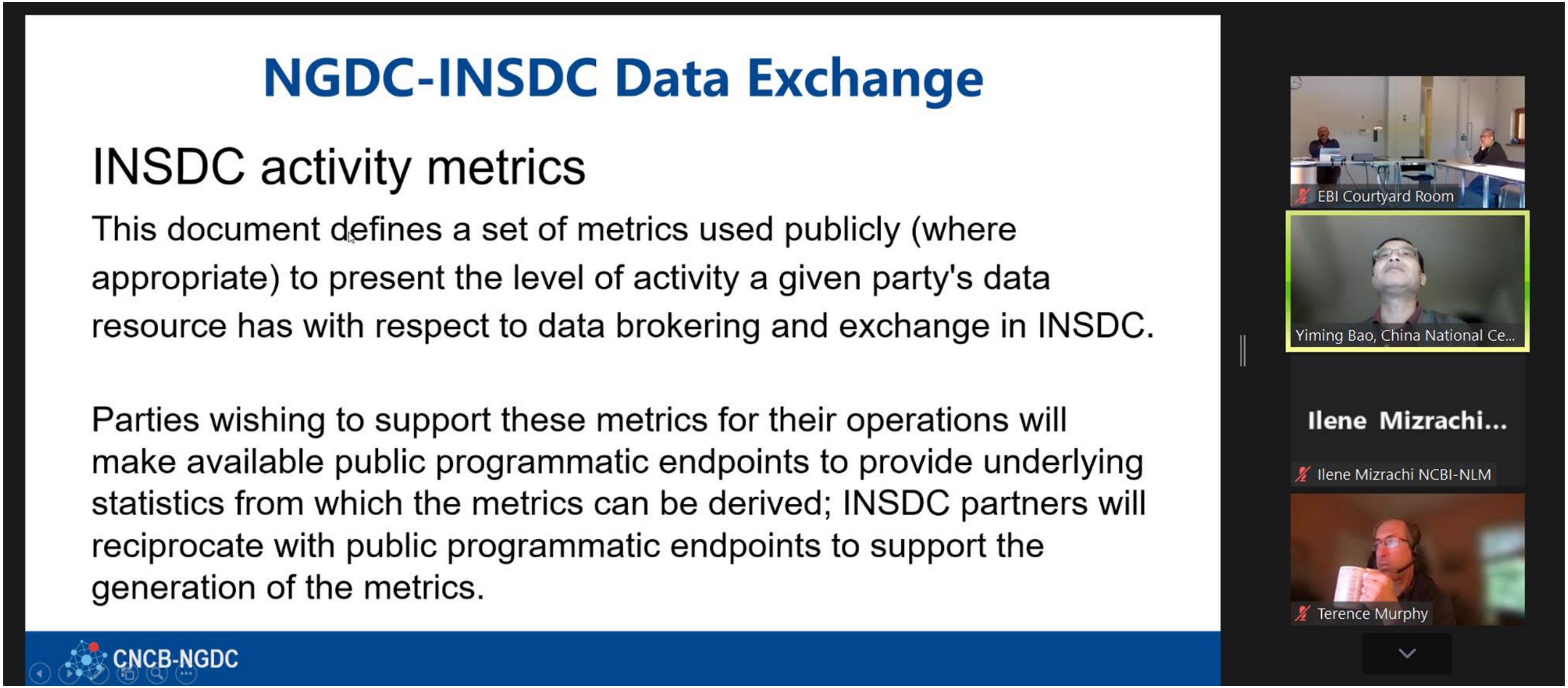
KEY DATES

21st March 2022 Submission of expressions of interest opened.
22nd April 2022 Deadline for submission of expressions of interest.
8th August 2022 Deadline for Submission of full applications.
Week commencing 12th December 2022 Announcement of the initial GCBR list.

GLOBAL CORE BIODATA RESOURCE SELECTION

https://globalbiodata.org/scientific-activities/gcbr-selection/

INSDC meeting



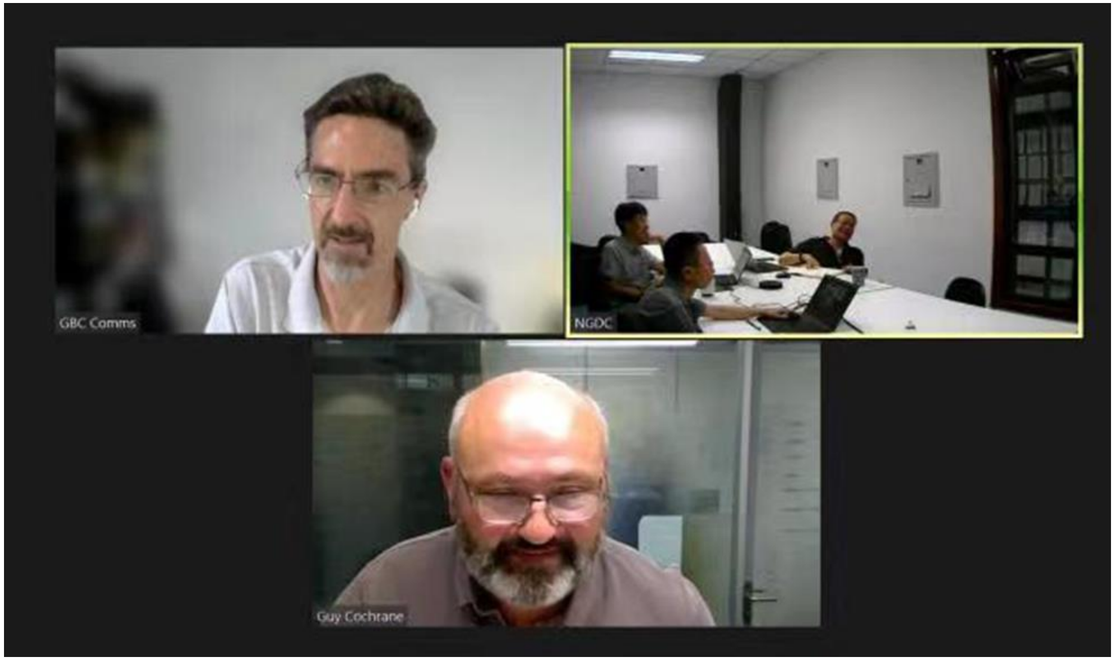
NGDC-INSDC Data Exchange

INSDC activity metrics

This document defines a set of metrics used publicly (where appropriate) to present the level of activity a given party's data resource has with respect to data brokering and exchange in INSDC.

Parties wishing to support these metrics for their operations will make available public programmatic endpoints to provide underlying statistics from which the metrics can be derived; INSDC partners will reciprocate with public programmatic endpoints to support the generation of the metrics.

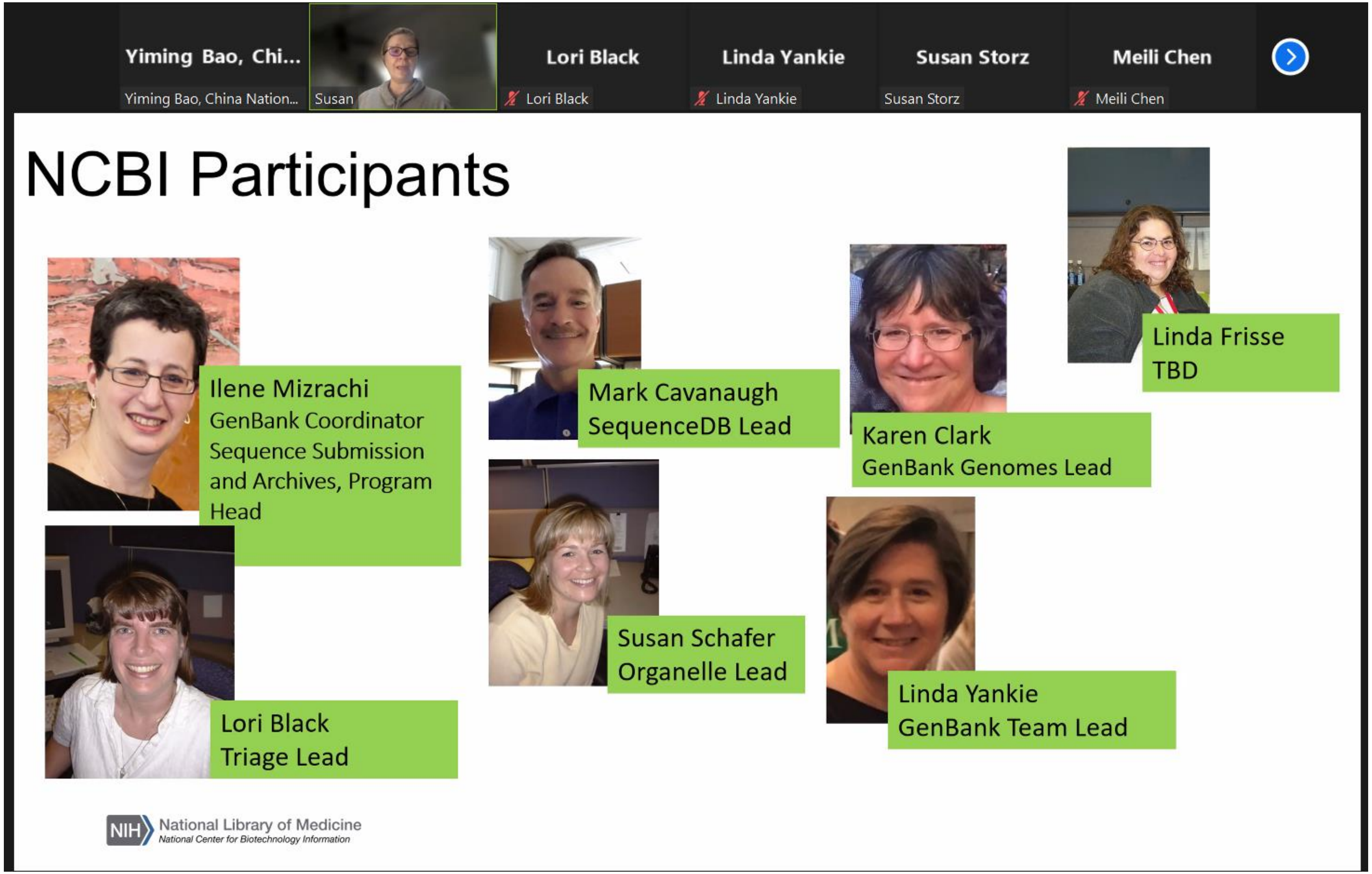
NGDC – EBI Discussions



Database Commons

Search Engine

NCBI Curation Training – 7X



NCBI Participants

Ilene Mizrachi
GenBank Coordinator
Sequence Submission and Archives, Program Head

Mark Cavanaugh
SequenceDB Lead

Karen Clark
GenBank Genomes Lead

Linda Frisse
TBD

Lori Black
Triage Lead

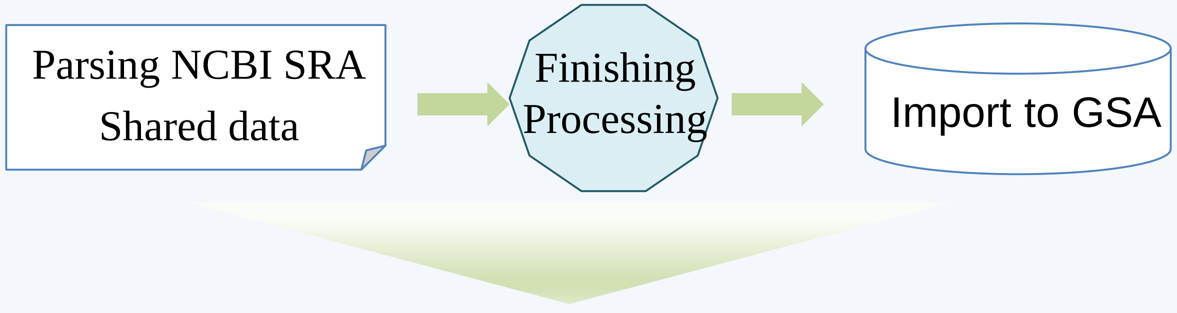
Susan Schafer
Organelle Lead

Linda Yankie
GenBank Team Lead

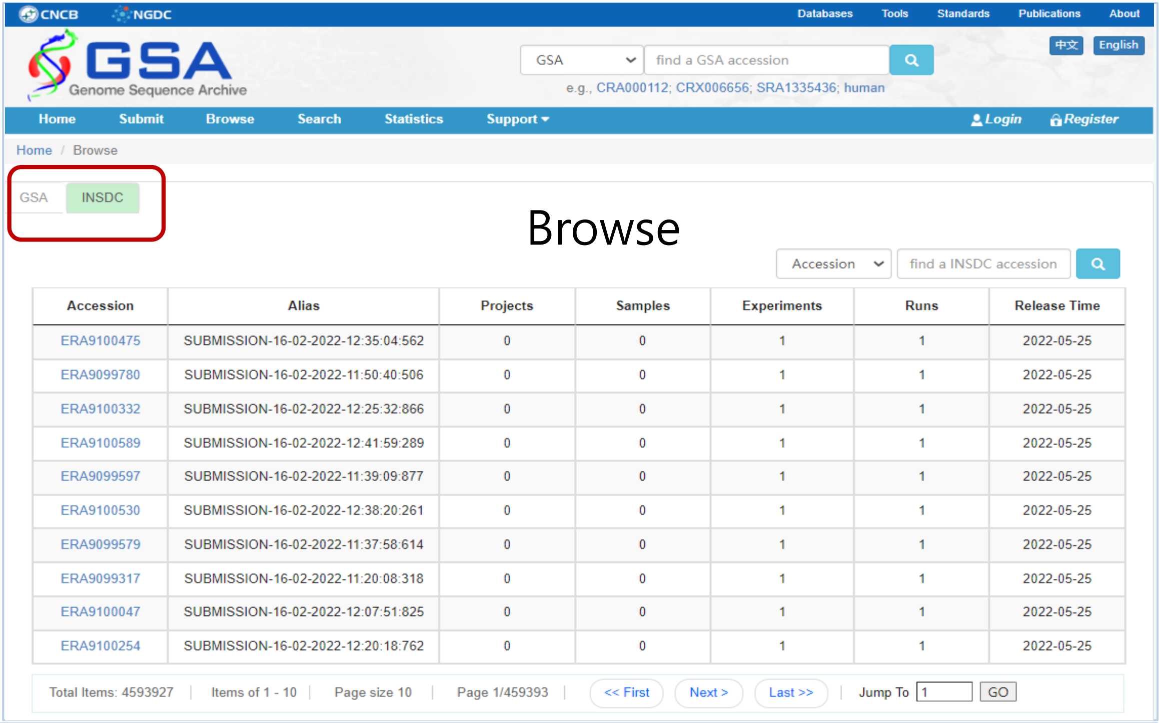
NIH National Library of Medicine
National Center for Biotechnology Information

INSDC Data Mirroring

BioSample, BioProject and SRA Data



623,345	29,562,590	21,994,135	23,523,117
PROJECTS	SAMPLES	EXPERIMENTS	RUNS



Browse

Accession | find a INSDC accession

Accession	Alias	Projects	Samples	Experiments	Runs	Release Time
ERA9100475	SUBMISSION-16-02-2022-12:35:04:562	0	0	1	1	2022-05-25
ERA9099780	SUBMISSION-16-02-2022-11:50:40:506	0	0	1	1	2022-05-25
ERA9100332	SUBMISSION-16-02-2022-12:25:32:866	0	0	1	1	2022-05-25
ERA9100589	SUBMISSION-16-02-2022-12:41:59:289	0	0	1	1	2022-05-25
ERA9099597	SUBMISSION-16-02-2022-11:39:09:877	0	0	1	1	2022-05-25
ERA9100530	SUBMISSION-16-02-2022-12:38:20:261	0	0	1	1	2022-05-25
ERA9099579	SUBMISSION-16-02-2022-11:37:58:614	0	0	1	1	2022-05-25
ERA9099317	SUBMISSION-16-02-2022-11:20:08:318	0	0	1	1	2022-05-25
ERA9100047	SUBMISSION-16-02-2022-12:07:51:825	0	0	1	1	2022-05-25
ERA9100254	SUBMISSION-16-02-2022-12:20:18:762	0	0	1	1	2022-05-25

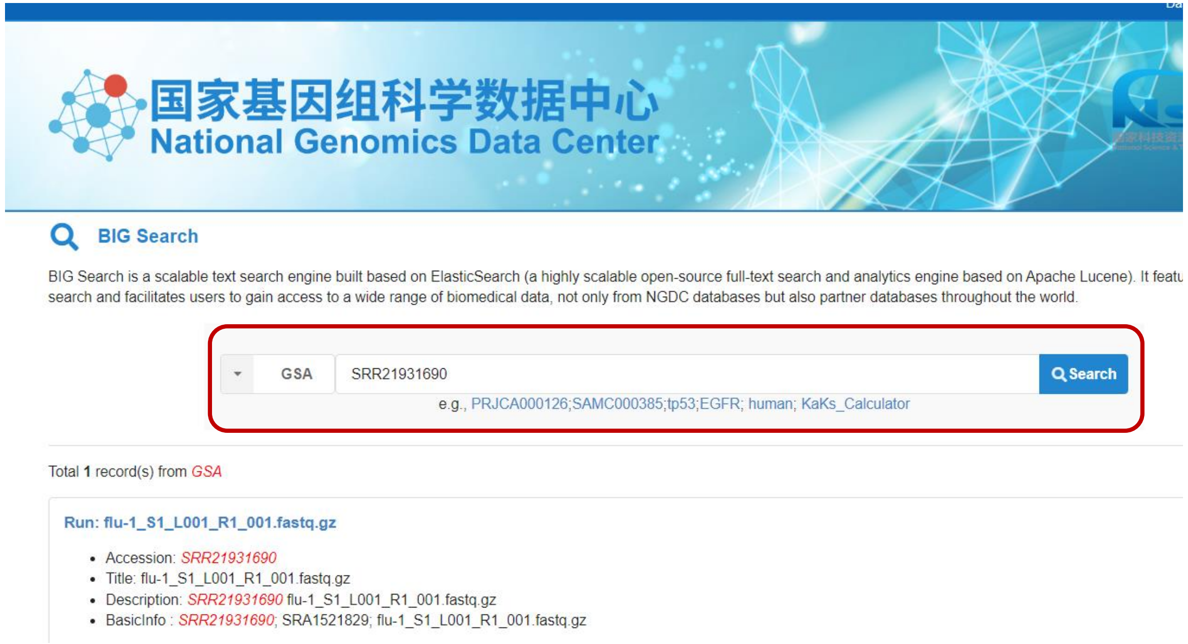
Total Items: 4593927 | Items of 1 - 10 | Page size 10 | Page 1/459393 | << First | Next > | Last >> | Jump To 1 GO

Metadata information has been updated to **2022-11-08**

The data files have been downloaded every day since **2022-04-20**

Data Files： **1.6 PB**

Integration of SRA Data



BIG Search

BIG Search is a scalable text search engine built based on ElasticSearch (a highly scalable open-source full-text search and analytics engine based on Apache Lucene). It featu search and facilitates users to gain access to a wide range of biomedical data, not only from NGDC databases but also partner databases throughout the world.

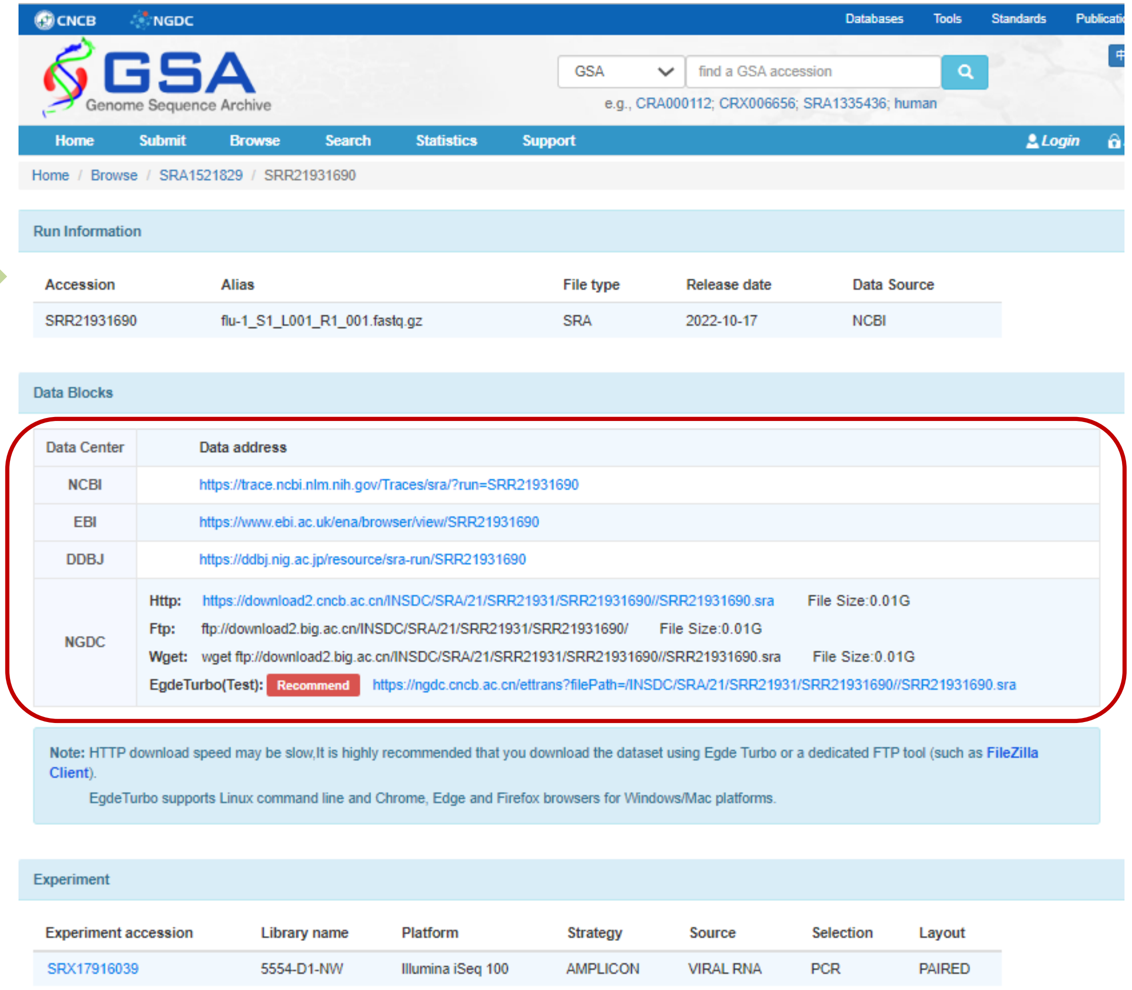
GSA | SRR21931690 | Search

e.g., PRJCA000126;SAMC000385;tp53;EGFR; human; KaKs_Calculator

Total **1** record(s) from *GSA*

Run: flu-1_S1_L001_R1_001.fastq.gz

- Accession: *SRR21931690*
- Title: flu-1_S1_L001_R1_001.fastq.gz
- Description: *SRR21931690* flu-1_S1_L001_R1_001.fastq.gz
- BasicInfo : *SRR21931690*; SRA1521829; flu-1_S1_L001_R1_001.fastq.gz



e.g., CRA000112; CRX006656; SRA1335436; human

Home / Browse / SRA1521829 / SRR21931690

Run Information

Accession	Alias	File type	Release date	Data Source
SRR21931690	flu-1_S1_L001_R1_001.fastq.gz	SRA	2022-10-17	NCBI

Data Blocks

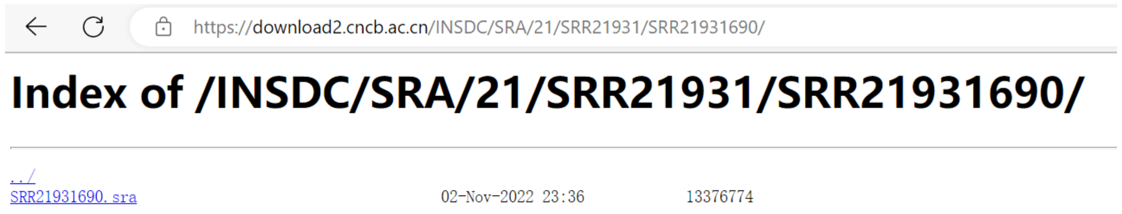
Data Center	Data address
NCBI	https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR21931690
EBI	https://www.ebi.ac.uk/ena/browser/view/SRR21931690
DDBJ	https://ddbj.nig.ac.jp/resource/sra-run/SRR21931690
NGDC	**Http:** https://download2.cncb.ac.cn/INSDC/SRA/21/SRR21931/SRR21931690//SRR21931690.sra File Size:0.01G **Ftp:** ftp://download2.big.ac.cn/INSDC/SRA/21/SRR21931/SRR21931690/ File Size:0.01G **Wget:** wget ftp://download2.big.ac.cn/INSDC/SRA/21/SRR21931/SRR21931690//SRR21931690.sra File Size:0.01G **EgdeTurbo(Test):** Recommend https://ngdc.cncb.ac.cn/ettrans?filePath=/INSDC/SRA/21/SRR21931/SRR21931690//SRR21931690.sra

Note: HTTP download speed may be slow,It is highly recommended that you download the dataset using Egde Turbo or a dedicated FTP tool (such as **FileZilla Client**).

EgdeTurbo supports Linux command line and Chrome, Edge and Firefox browsers for Windows/Mac platforms.

Experiment

Experiment accession	Library name	Platform	Strategy	Source	Selection	Layout
SRX17916039	5554-D1-NW	Illumina iSeq 100	AMPLICON	VIRAL RNA	PCR	PAIRED



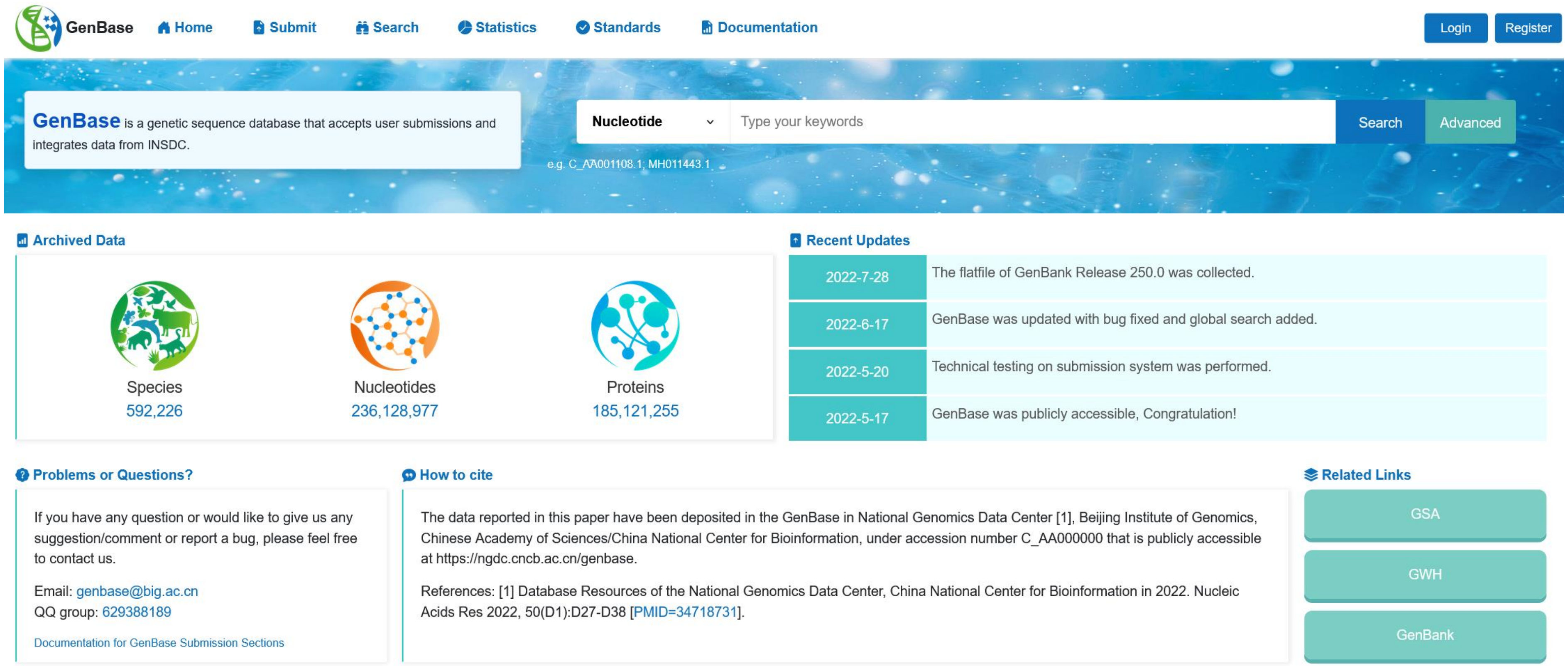
https://download2.cncb.ac.cn/INSDC/SRA/21/SRR21931/SRR21931690/

Index of /INSDC/SRA/21/SRR21931/SRR21931690/

```
../
SRR21931690.sra                                    02-Nov-2022 23:36          13376774
```

https://download2.cncb.ac.cn/INSDC/SRA/21/SRR21931/SRR21931690/

GenBase



GenBase is a genetic sequence database that accepts user submissions and integrates data from INSDC.

Archived Data

Species	Nucleotides	Proteins
592,226	236,128,977	185,121,255

Recent Updates

2022-7-28	The flatfile of GenBank Release 250.0 was collected.
2022-6-17	GenBase was updated with bug fixed and global search added.
2022-5-20	Technical testing on submission system was performed.
2022-5-17	GenBase was publicly accessible, Congratulation!

Problems or Questions?

If you have any question or would like to give us any suggestion/comment or report a bug, please feel free to contact us.

Email: genbase@big.ac.cn
QQ group: 629388189

Documentation for GenBase Submission Sections

How to cite

The data reported in this paper have been deposited in the GenBase in National Genomics Data Center [1], Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under accession number C_AA000000 that is publicly accessible at https://ngdc.cncb.ac.cn/genbase.

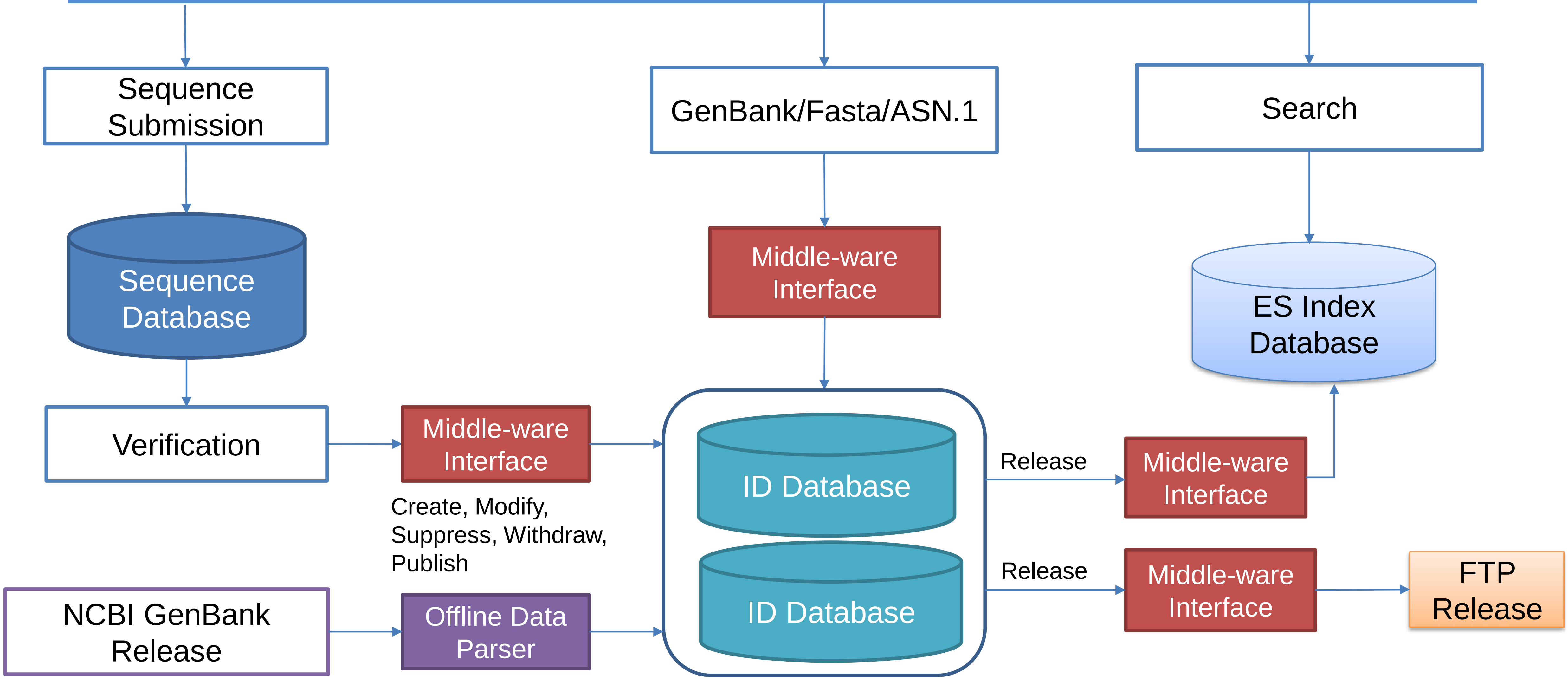
References: [1] Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2022. Nucleic Acids Res 2022, 50(D1):D27-D38 [PMID=34718731].

Related Links

- **GenBank/ENA** type database, **GenBank Release 250.0** has been integrated
- **In total**: **592,226** Species, **236,128,977** Nucleotides, **185,121,255** Proteins
- **Direct submission**: **1300** Nucleotides, **1073** Proteins

https://ngdc.cncb.ac.cn/genbase/

GenBase System



Sequence Submission

Sequence Database

Verification

Middle-ware Interface

Create, Modify, Suppress, Withdraw, Publish

NCBI GenBank Release

Offline Data Parser

GenBank/Fasta/ASN.1

Middle-ware Interface

ID Database

ID Database

Release

Release

Search

ES Index Database

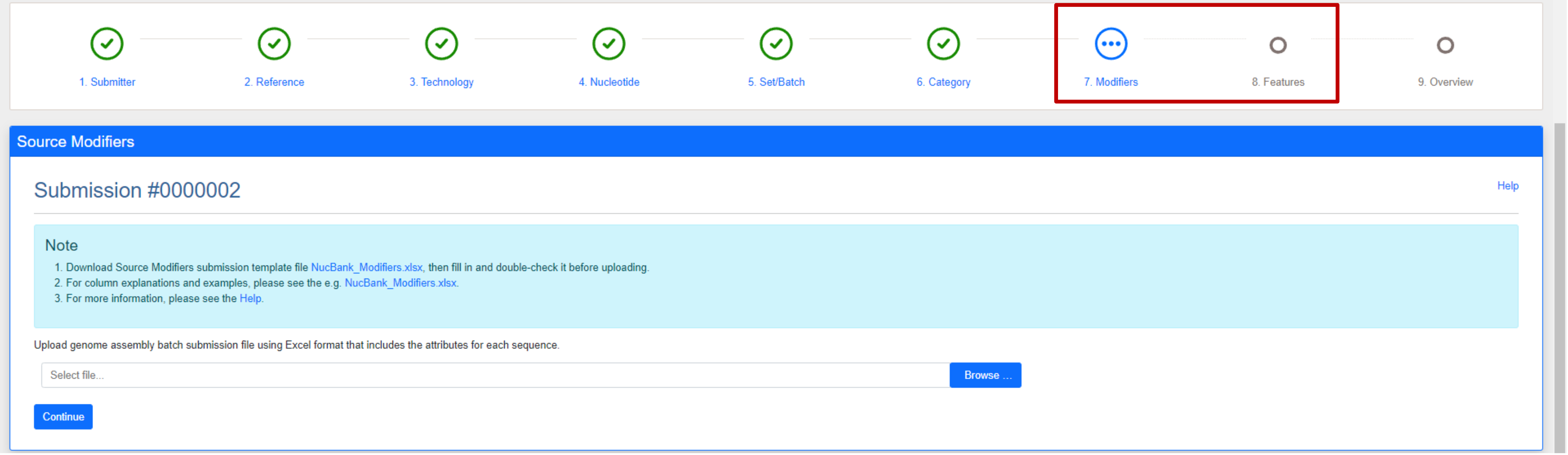
Middle-ware Interface

Middle-ware Interface

FTP Release

Submission Portal

Source Modifiers Table   Feature table



Source Modifiers

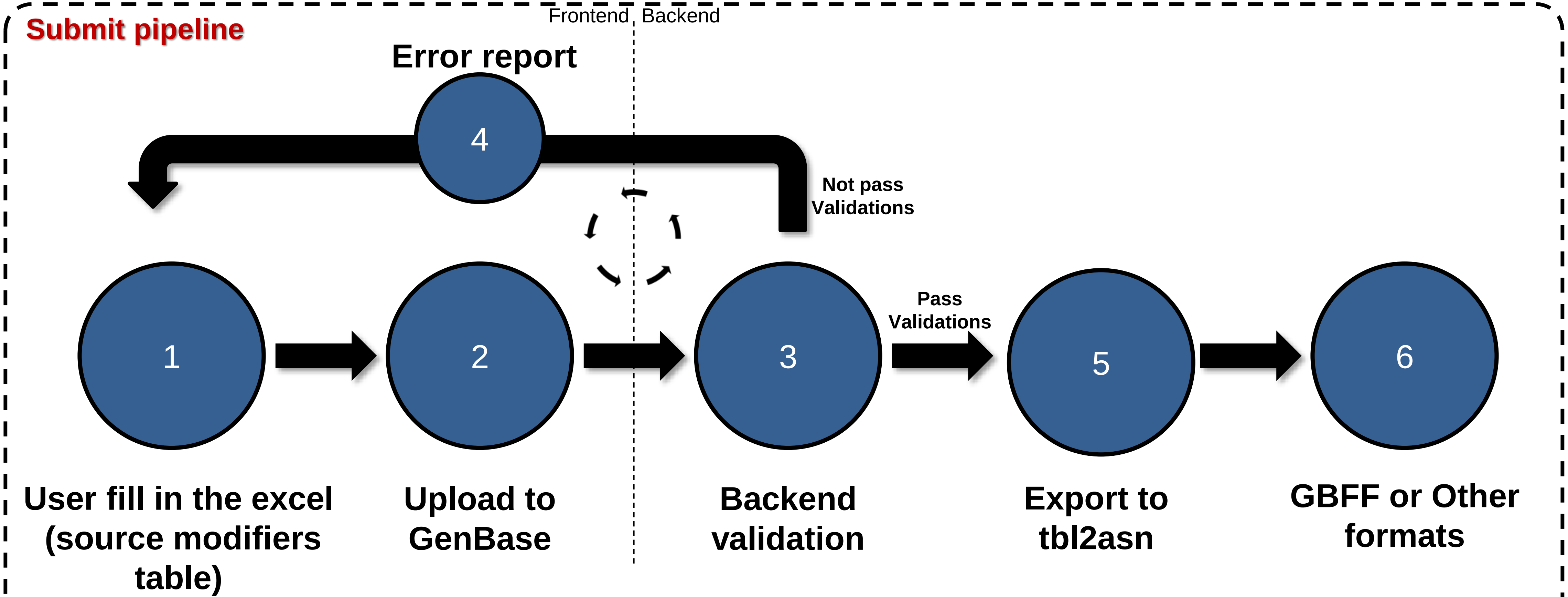
Submission #0000002

Note

1. Download Source Modifiers submission template file NucBank_Modifiers.xlsx, then fill in and double-check it before uploading.
2. For column explanations and examples, please see the e.g. NucBank_Modifiers.xlsx.
3. For more information, please see the Help.

Upload genome assembly batch submission file using Excel format that includes the attributes for each sequence.

Source Modifiers Table



Submit pipeline

Frontend

Backend

Error report

Not pass Validations

Pass Validations

1 User fill in the excel (source modifiers table)

2 Upload to GenBase

3 Backend validation

4

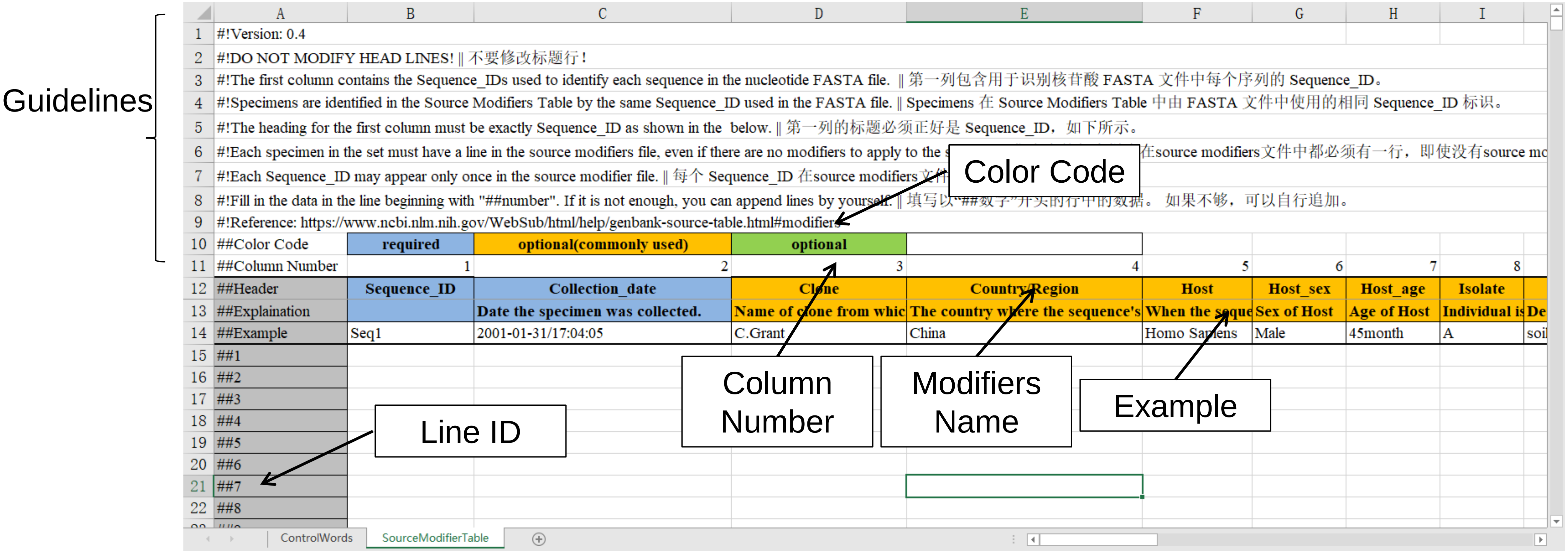
5 Export to tbl2asn

6 GBFF or Other formats

- Source modifiers are compatible with NCBI source modifiers[1,2])

1. https://ngdc.cncb.ac.cn/genbase/standards
2. https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers

Excel structure



1 #!Version: 0.4
2 #!DO NOT MODIFY HEAD LINES! || 不要修改标题行！
3 #!The first column contains the Sequence_IDs used to identify each sequence in the nucleotide FASTA file. || 第一列包含用于识别核苷酸 FASTA 文件中每个序列的 Sequence_ID。
4 #!Specimens are identified in the Source Modifiers Table by the same Sequence_ID used in the FASTA file. || Specimens 在 Source Modifiers Table 中由 FASTA 文件中使用的相同 Sequence_ID 标识。
5 #!The heading for the first column must be exactly Sequence_ID as shown in the below. || 第一列的标题必须正好是 Sequence_ID，如下所示。
6 #!Each specimen in the set must have a line in the source modifiers file, even if there are no modifiers to apply to the s 在source modifiers文件中都必须有一行，即使没有source mo
7 #!Each Sequence_ID may appear only once in the source modifier file. || 每个 Sequence_ID 在source modifiers文件
8 #!Fill in the data in the line beginning with "##number". If it is not enough, you can append lines by yourself. || 填写以“##数字”开头的行中的数据。 如果不够，可以自行追加。
9 #!Reference: https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers

	A	B	C	D	E	F	G	H	I	
10	##Color Code	required	optional(commonly used)	optional						
11	##Column Number	1	2	3	4	5	6	7	8	
12	##Header	Sequence_ID	Collection_date	Clone	Country/Region	Host	Host_sex	Host_age	Isolate	
13	##Explaination		Date the specimen was collected.	Name of clone from whic	The country where the sequence's	When the seque	Sex of Host	Age of Host	Individual is	De
14	##Example	Seq1	2001-01-31/17:04:05	C.Grant	China	Homo Sapiens	Male	45month	A	soi
15	##1									
16	##2									
17	##3									
18	##4									
19	##5									
20	##6									
21	##7									
22	##8									

- Guidelines describe simple help messages, both English and Chinese messages are provided
- Color Code mark each modifiers, and classify them to required, optional (recommended), optional
- Each cell is identified by Line ID and Column number, so that users can focus on them faster

Validators

Structure Validators

Validate table structure for parsing data correctly.

- Line ID validator: Line ID in the first column is required, if not , fatal error will be thrown.

Column Validators

Validate data by column.

- Column_unique_value_validator: columns with duplicate values are not allowed

Row Validators

Validate data by row.

- Row_co_occur_validator: check whether specified values are co-occurred
- Row_mutex_validator: specified values are mutually exclusive

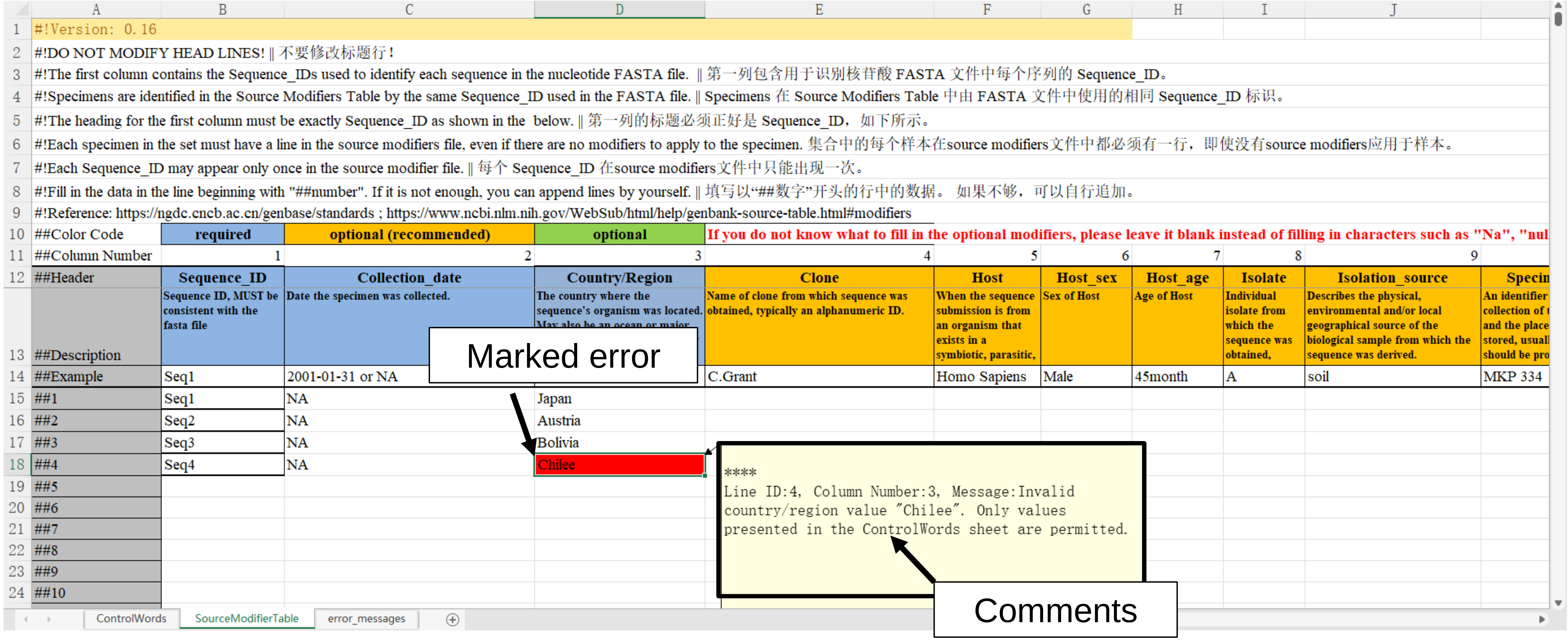
Cell Validators

Same validators are applied to all cells in one column. If errors are detected, the cell will be mark red, and comment will be add.

- Control vocabulary validators
- Format validators

(sourceModifiersValidator.py)

Validation Example



#!Version: 0. 16

#!DO NOT MODIFY HEAD LINES! ‖ 不要修改标题行！

#!The first column contains the Sequence_IDs used to identify each sequence in the nucleotide FASTA file. ‖ 第一列包含用于识别核苷酸 FASTA 文件中每个序列的 Sequence_ID。

#!Specimens are identified in the Source Modifiers Table by the same Sequence_ID used in the FASTA file. ‖ Specimens 在 Source Modifiers Table 中由 FASTA 文件中使用的相同 Sequence_ID 标识。

#!The heading for the first column must be exactly Sequence_ID as shown in the below. ‖ 第一列的标题必须正好是 Sequence_ID，如下所示。

#!Each specimen in the set must have a line in the source modifiers file, even if there are no modifiers to apply to the specimen. 集合中的每个样本在source modifiers文件中都必须有一行，即使没有source modifiers应用于样本。

#!Each Sequence_ID may appear only once in the source modifier file. ‖ 每个 Sequence_ID 在source modifiers文件中只能出现一次。

#!Fill in the data in the line beginning with "##number". If it is not enough, you can append lines by yourself. ‖ 填写以“##数字”开头的行中的数据。 如果不够，可以自行追加。

#!Reference: https://ngdc.cncb.ac.cn/genbase/standards ; https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers

A	B	C	D	E	F	G	H	I	J	
##Color Code	required	optional (recommended)	optional	If you do not know what to fill in the optional modifiers, please leave it blank instead of filling in characters such as "Na", "nul						
##Column Number	1	2	3	4	5	6	7	8	9	
##Header	Sequence_ID	Collection_date	Country/Region	Clone	Host	Host_sex	Host_age	Isolate	Isolation_source	Specin
##Description	Sequence ID, MUST be consistent with the fasta file	Date the specimen was collected.	The country where the sequence's organism was located. May also be an ocean or major	Name of clone from which sequence was obtained, typically an alphanumeric ID.	When the sequence submission is from an organism that exists in a symbiotic, parasitic,	Sex of Host	Age of Host	Individual isolate from which the sequence was obtained,	Describes the physical, environmental and/or local geographical source of the biological sample from which the sequence was derived.	An identifier collection of t and the place stored, usuall should be pro
##Example	Seq1	2001-01-31 or NA		C.Grant	Homo Sapiens	Male	45month	A	soil	MKP 334
##1	Seq1	NA	Japan							
##2	Seq2	NA	Austria							
##3	Seq3	NA	Bolivia							
##4	Seq4	NA	Chilee							
##5										
##6										
##7										
##8										
##9										
##10										

****
Line ID:4, Column Number:3, Message:Invalid country/region value "Chilee". Only values presented in the ControlWords sheet are permitted.

Processed by **sourceModifiersValidator.py**

Feature Table

Three optional formats for submitting sequence features

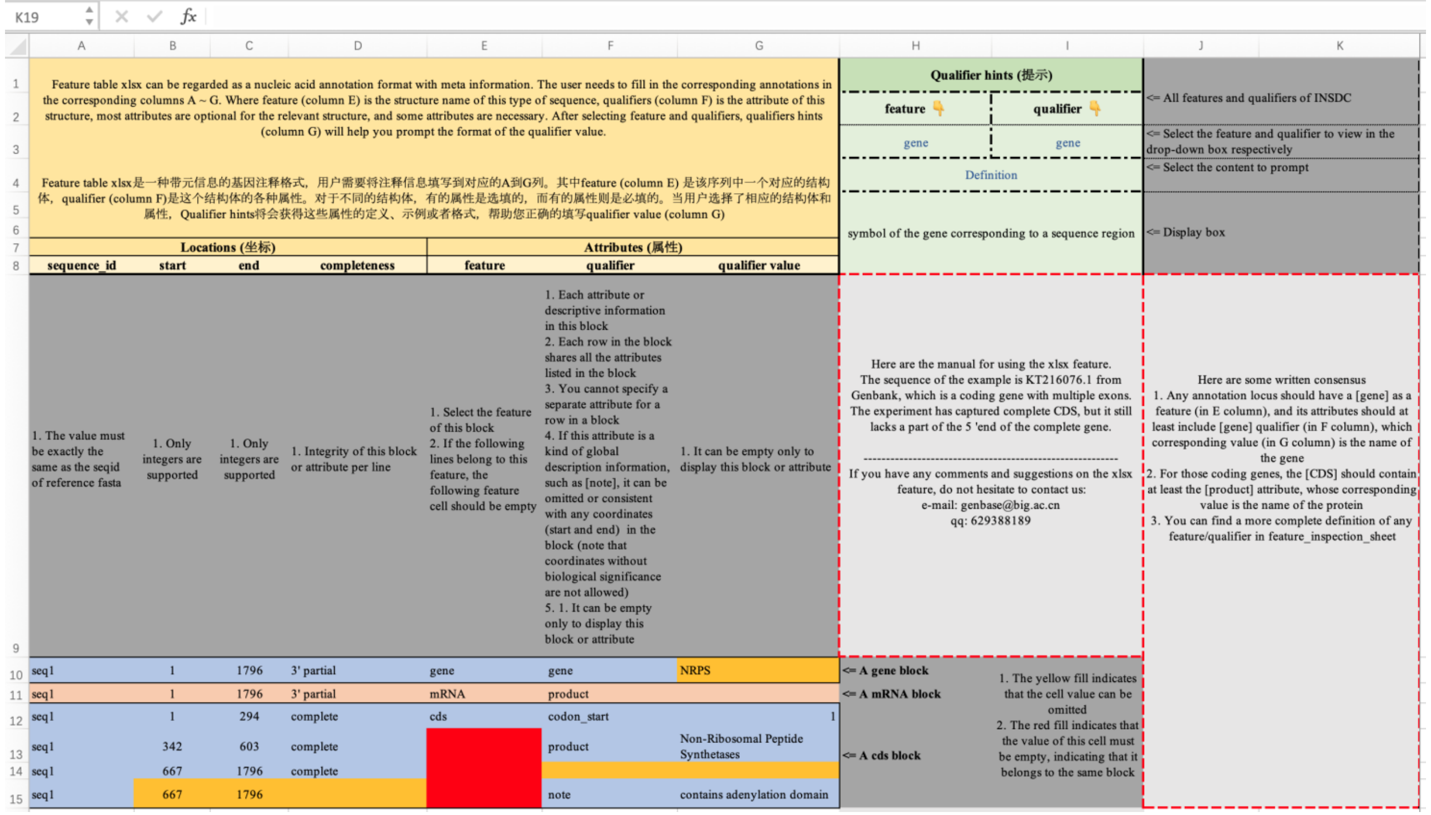
Plain text tbl or gff3: >10 sequences

```
>Feature Seq1
<1    >1050    gene
                        gene          ATH1
<1    1009    CDS
                        product       acid trehalase
                        product       Athlp
                        codon_start   2
<1    >1050    mRNA
                        product       acid trehalase

>Feature Seq2
2626  2590    tRNA
2570  2535
                        product       tRNA-Phe

>Feature Seq3
1080  1210  CDS
1275  1315
                        product       actin
                        note          alternatively spliced
1055  1210  mRNA
1275  1340
                        product       actin
1055  1340  gene
                        gene          ACT
1055  1079  5'UTR
1316  1340  3'UTR
```

Excel tbl (with hints): <10 sequences



K19

Feature table xlsx can be regarded as a nucleic acid annotation format with meta information. The user needs to fill in the corresponding annotations in the corresponding columns A ~ G. Where feature (column E) is the structure name of this type of sequence, qualifiers (column F) is the attribute of this structure, most attributes are optional for the relevant structure, and some attributes are necessary. After selecting feature and qualifiers, qualifiers hints (column G) will help you prompt the format of the qualifier value.

Feature table xlsx是一种带元信息的基因注释格式，用户需要将注释信息填写到对应的A到G列。其中feature (column E) 是该序列中一个对应的结构体，qualifier (column F)是这个结构体的各种属性。对于不同的结构体，有的属性是选填的，而有的属性则是必填的。当用户选择了相应的结构体和属性，Qualifier hints将会获得这些属性的定义、示例或者格式，帮助您正确的填写qualifier value (column G)

Locations (坐标)				Attributes (属性)		
sequence_id	start	end	completeness	feature	qualifier	qualifier value
1. The value must be exactly the same as the seqid of reference fasta	1. Only integers are supported	1. Only integers are supported	1. Integrity of this block or attribute per line	1. Select the feature of this block 2. If the following lines belong to this feature, the following feature cell should be empty	1. Each attribute or descriptive information in this block 2. Each row in the block shares all the attributes listed in the block 3. You cannot specify a separate attribute for a row in a block 4. If this attribute is a kind of global description information, such as [note], it can be omitted or consistent with any coordinates (start and end) in the block (note that coordinates without biological significance are not allowed) 5. 1. It can be empty only to display this block or attribute	1. It can be empty only to display this block or attribute
seq1	1	1796	3' partial	gene	gene	NRPS
seq1	1	1796	3' partial	mRNA	product	
seq1	1	294	complete	cds	codon_start	1
seq1	342	603	complete	🟥	product	Non-Ribosomal Peptide Synthetases
seq1	667	1796	complete	🟥		
seq1	667	1796		🟥	note	contains adenylation domain

Qualifier hints (提示)

feature	qualifier
gene	gene

<= All features and qualifiers of INSDC

<= Select the feature and qualifier to view in the drop-down box respectively

Definition

<= Select the content to prompt

symbol of the gene corresponding to a sequence region

<= Display box

Here are the manual for using the xlsx feature.
The sequence of the example is KT216076.1 from Genbank, which is a coding gene with multiple exons. The experiment has captured complete CDS, but it still lacks a part of the 5 'end of the complete gene.

If you have any comments and suggestions on the xlsx feature, do not hesitate to contact us:
e-mail: genbase@big.ac.cn
qq: 629388189

Here are some written consensus
1. Any annotation locus should have a [gene] as a feature (in E column), and its attributes should at least include [gene] qualifier (in F column), which corresponding value (in G column) is the name of the gene
2. For those coding genes, the [CDS] should contain at least the [product] attribute, whose corresponding value is the name of the protein
3. You can find a more complete definition of any feature/qualifier in feature_inspection_sheet

<= A gene block

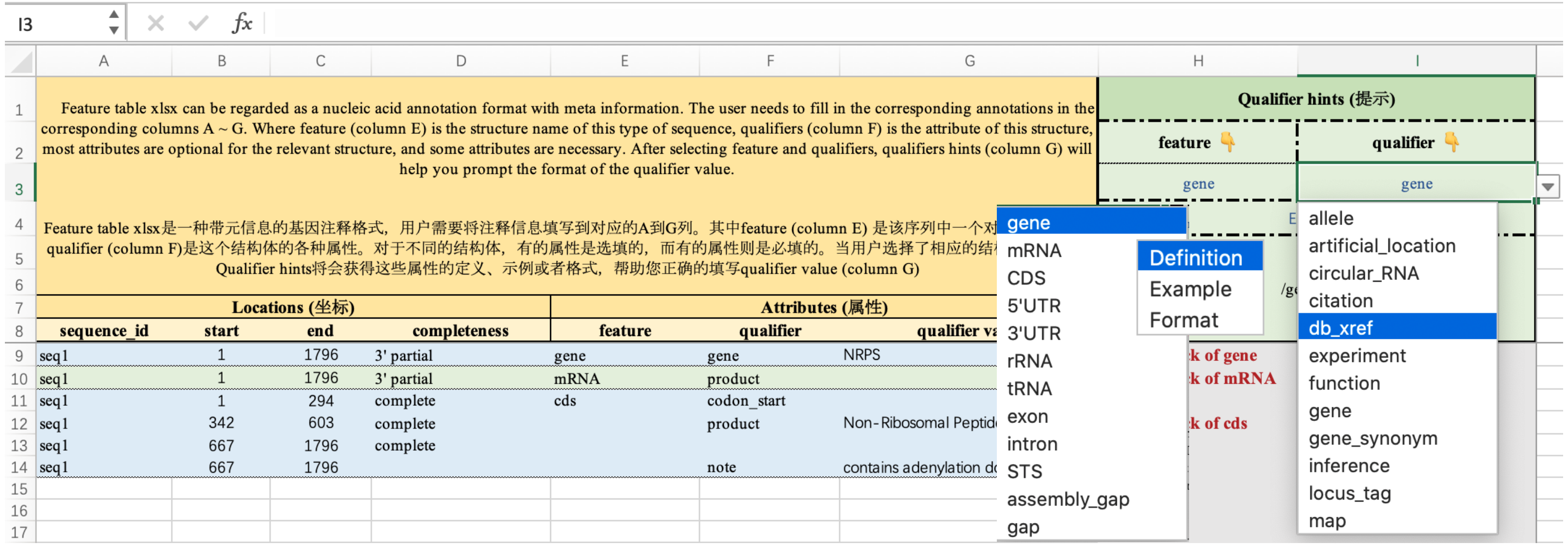
<= A mRNA block

<= A cds block

1. The yellow fill indicates that the cell value can be omitted
2. The red fill indicates that the value of this cell must be empty, indicating that it belongs to the same block

CNCB-NGDC

Feature table with INSDC standards



Feature table xlsx can be regarded as a nucleic acid annotation format with meta information. The user needs to fill in the corresponding annotations in the corresponding columns A ~ G. Where feature (column E) is the structure name of this type of sequence, qualifiers (column F) is the attribute of this structure, most attributes are optional for the relevant structure, and some attributes are necessary. After selecting feature and qualifiers, qualifiers hints (column G) will help you prompt the format of the qualifier value.

Feature table xlsx是一种带元信息的基因注释格式，用户需要将注释信息填写到对应的A到G列。其中feature (column E) 是该序列中一个对
qualifier (column F)是这个结构体的各种属性。对于不同的结构体，有的属性是选填的，而有的属性则是必填的。当用户选择了相应的结构
Qualifier hints将会获得这些属性的定义、示例或者格式，帮助您正确的填写qualifier value (column G)

Locations (坐标)				Attributes (属性)		
sequence_id	start	end	completeness	feature	qualifier	qualifier va
seq1	1	1796	3' partial	gene	gene	NRPS
seq1	1	1796	3' partial	mRNA	product	
seq1	1	294	complete	cds	codon_start	
seq1	342	603	complete		product	Non-Ribosomal Peptid
seq1	667	1796	complete			
seq1	667	1796			note	contains adenylation d

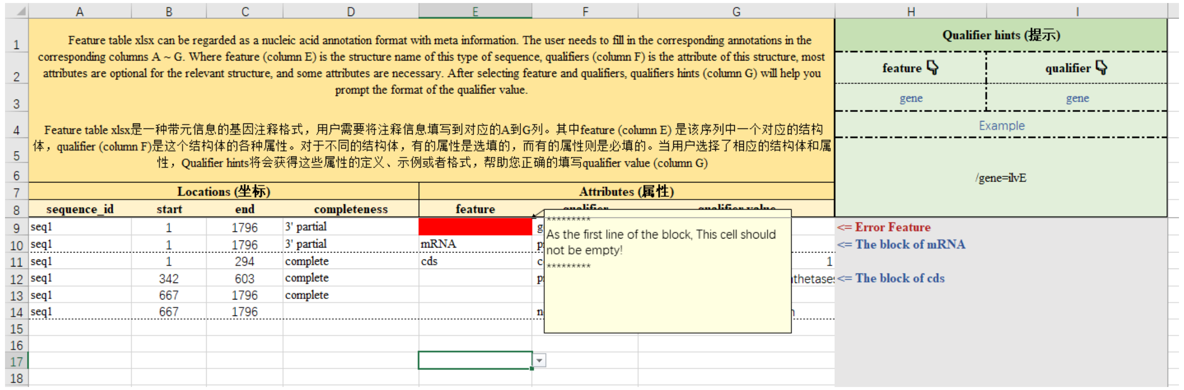
Qualifier hints (提示)

feature	qualifier
gene	gene

k of gene
k of mRNA
k of cds

We embed the INSDC standard into excel, and assist users to fill in the qualifier value by prompting for *formats*, *definitions* and *examples*.

Different features correspond to their respective qualifiers, and also correspond to different hints.



Feature table xlsx can be regarded as a nucleic acid annotation format with meta information. The user needs to fill in the corresponding annotations in the corresponding columns A ~ G. Where feature (column E) is the structure name of this type of sequence, qualifiers (column F) is the attribute of this structure, most attributes are optional for the relevant structure, and some attributes are necessary. After selecting feature and qualifiers, qualifiers hints (column G) will help you prompt the format of the qualifier value.

Feature table xlsx是一种带元信息的基因注释格式，用户需要将注释信息填写到对应的A到G列。其中feature (column E) 是该序列中一个对应的结构体，qualifier (column F)是这个结构体的各种属性。对于不同的结构体，有的属性是选填的，而有的属性则是必填的。当用户选择了相应的结构体和属性，Qualifier hints将会获得这些属性的定义、示例或者格式，帮助您正确的填写qualifier value (column G)

Locations (坐标)				Attributes (属性)		
sequence_id	start	end	completeness	feature	qualifier	qualifier value
seq1	1	1796	3' partial		g	
seq1	1	1796	3' partial	mRNA	p	
seq1	1	294	complete	cds	c	1
seq1	342	603	complete		p	thetase
seq1	667	1796	complete			
seq1	667	1796			n	

****
As the first line of the block, This cell should not be empty!
****

<= Error Feature
<= The block of mRNA
<= The block of cds

Qualifier hints (提示)

feature	qualifier
gene	gene

Example

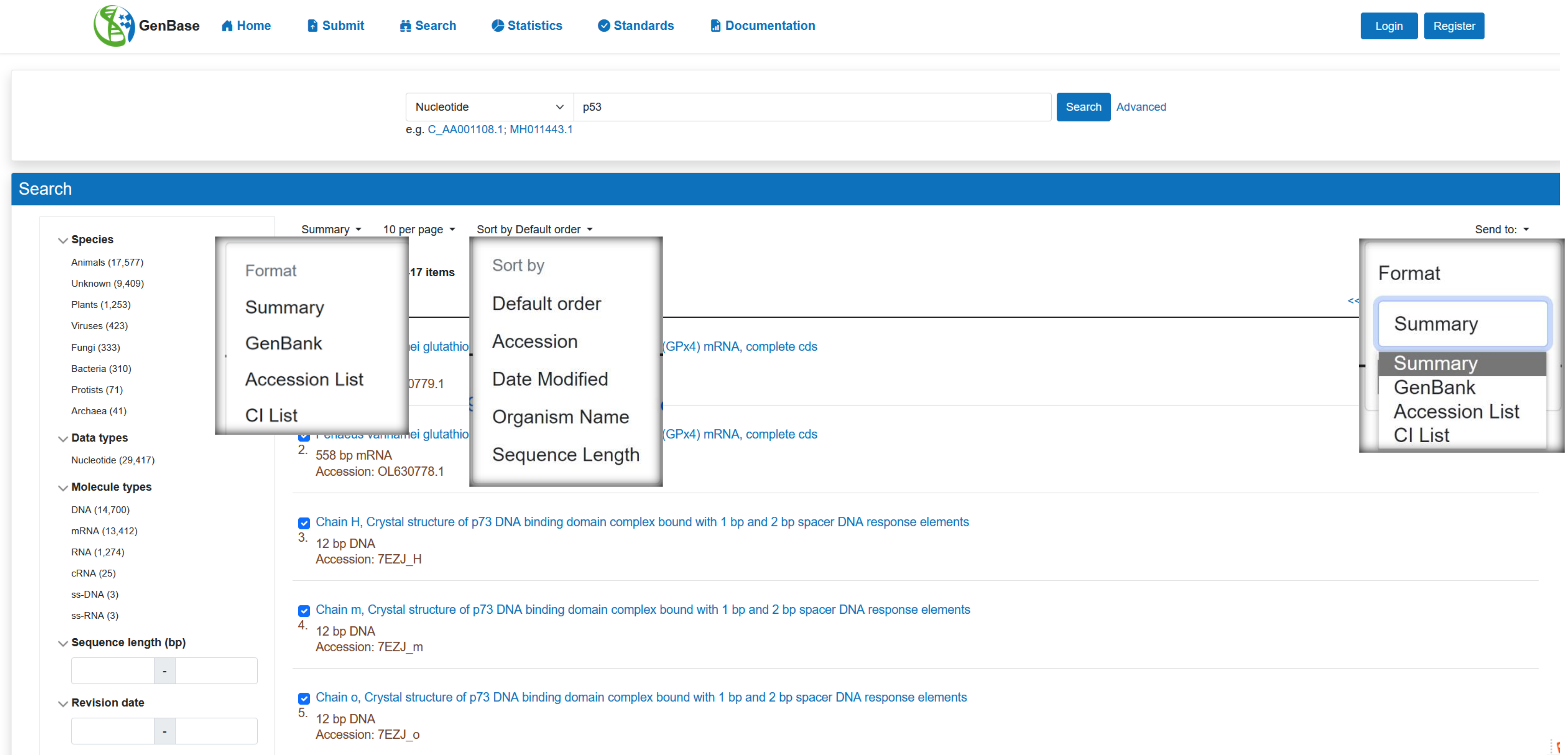
/gene=ilvE

We will feed back to user in real time the data that does not conform to the given format, and mark it in red.

User only need to re-submit the corrected excel again until the verification passes. Formatted tbl format will then be used for subsequent steps.

CNCB-NGDC

Search and Download



e.g. C_AA001108.1; MH011443.1

Species

- Animals (17,577)
- Unknown (9,409)
- Plants (1,253)
- Viruses (423)
- Fungi (333)
- Bacteria (310)
- Protists (71)
- Archaea (41)

Data types

- Nucleotide (29,417)

Molecule types

- DNA (14,700)
- mRNA (13,412)
- RNA (1,274)
- cRNA (25)
- ss-DNA (3)
- ss-RNA (3)

Sequence length (bp)

-

Revision date

-

17 items

ei glutathio (GPx4) mRNA, complete cds
0779.1

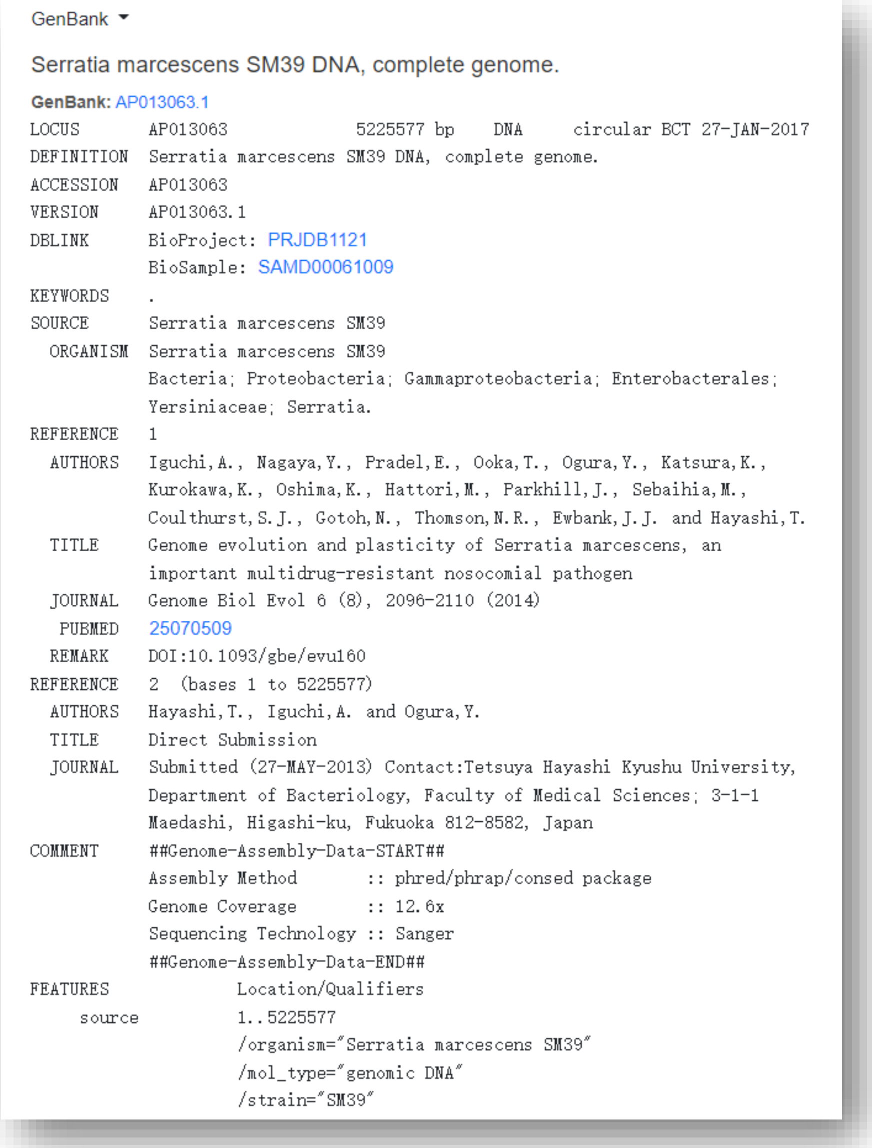
☑ Penaeus vannamei glutathio (GPx4) mRNA, complete cds
2. 558 bp mRNA
Accession: OL630778.1

☑ Chain H, Crystal structure of p73 DNA binding domain complex bound with 1 bp and 2 bp spacer DNA response elements
3. 12 bp DNA
Accession: 7EZJ_H

☑ Chain m, Crystal structure of p73 DNA binding domain complex bound with 1 bp and 2 bp spacer DNA response elements
4. 12 bp DNA
Accession: 7EZJ_m

☑ Chain o, Crystal structure of p73 DNA binding domain complex bound with 1 bp and 2 bp spacer DNA response elements
5. 12 bp DNA
Accession: 7EZJ_o

Display of Flat Files

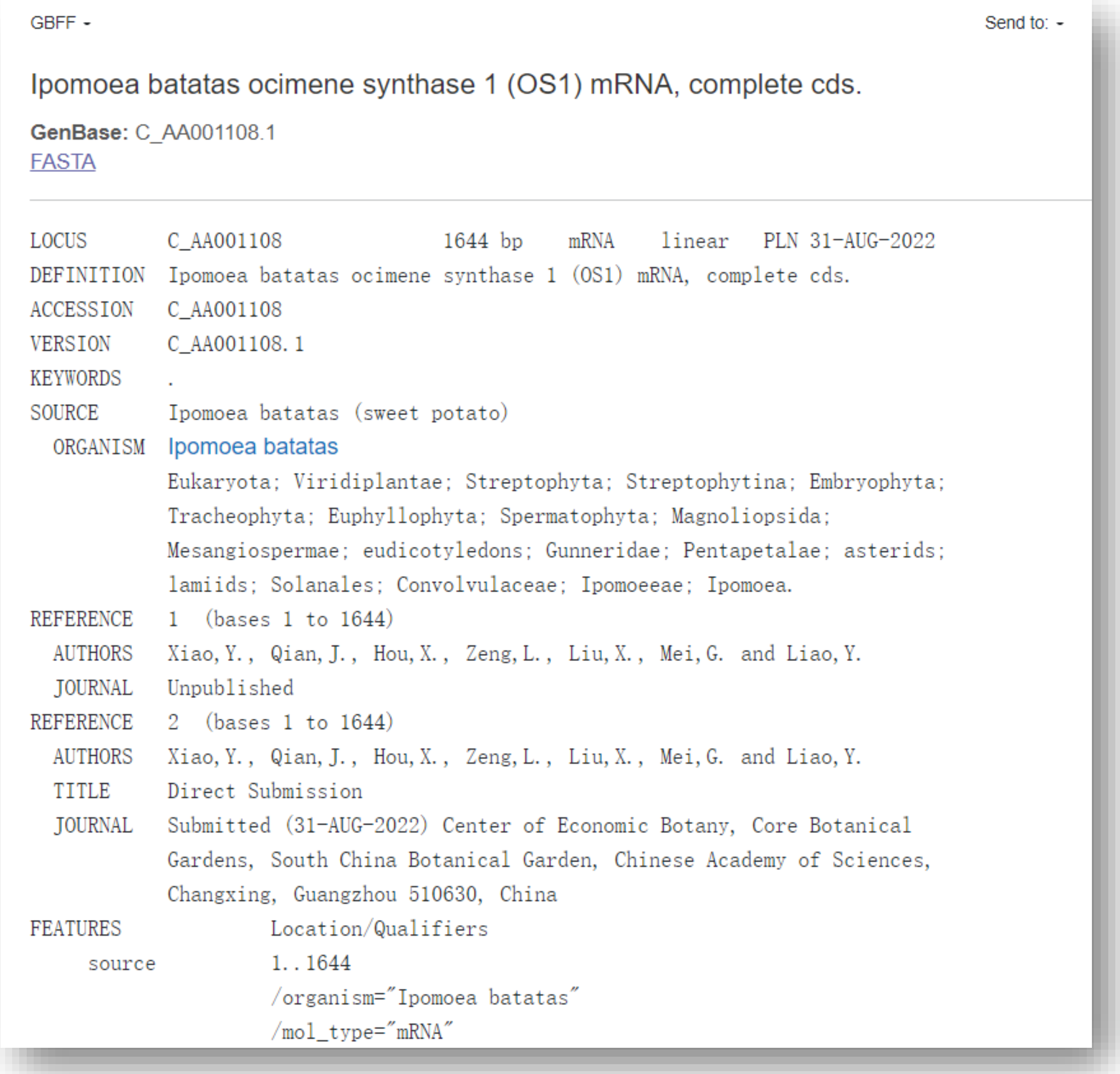


GenBank ▾

Serratia marcescens SM39 DNA, complete genome.

GenBank: AP013063.1

```
LOCUS       AP013063             5225577 bp    DNA     circular BCT 27-JAN-2017
DEFINITION  Serratia marcescens SM39 DNA, complete genome.
ACCESSION   AP013063
VERSION     AP013063.1
DBLINK      BioProject: PRJDB1121
            BioSample: SAMD00061009
KEYWORDS    .
SOURCE      Serratia marcescens SM39
  ORGANISM  Serratia marcescens SM39
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Yersiniaceae; Serratia.
REFERENCE   1
  AUTHORS   Iguchi,A., Nagaya,Y., Pradel,E., Ooka,T., Ogura,Y., Katsura,K.,
            Kurokawa,K., Oshima,K., Hattori,M., Parkhill,J., Sebaihia,M.,
            Coulthurst,S.J., Gotoh,N., Thomson,N.R., Ewbank,J.J. and Hayashi,T.
  TITLE     Genome evolution and plasticity of Serratia marcescens, an
            important multidrug-resistant nosocomial pathogen
  JOURNAL   Genome Biol Evol 6 (8), 2096-2110 (2014)
   PUBMED   25070509
  REMARK    DOI:10.1093/gbe/evu160
REFERENCE   2  (bases 1 to 5225577)
  AUTHORS   Hayashi,T., Iguchi,A. and Ogura,Y.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-MAY-2013) Contact:Tetsuya Hayashi Kyushu University,
            Department of Bacteriology, Faculty of Medical Sciences; 3-1-1
            Maedashi, Higashi-ku, Fukuoka 812-8582, Japan
COMMENT     ##Genome-Assembly-Data-START##
            Assembly Method       :: phred/phrap/consed package
            Genome Coverage       :: 12.6x
            Sequencing Technology :: Sanger
            ##Genome-Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..5225577
                     /organism="Serratia marcescens SM39"
                     /mol_type="genomic DNA"
                     /strain="SM39"
```



GBFF ▾ Send to: ▾

Ipomoea batatas ocimene synthase 1 (OS1) mRNA, complete cds.

GenBase: C_AA001108.1

FASTA

```
LOCUS       C_AA001108              1644 bp    mRNA    linear   PLN 31-AUG-2022
DEFINITION  Ipomoea batatas ocimene synthase 1 (OS1) mRNA, complete cds.
ACCESSION   C_AA001108
VERSION     C_AA001108.1
KEYWORDS    .
SOURCE      Ipomoea batatas (sweet potato)
  ORGANISM  Ipomoea batatas
            Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta;
            Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida;
            Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; asterids;
            lamiids; Solanales; Convolvulaceae; Ipomoeeae; Ipomoea.
REFERENCE   1  (bases 1 to 1644)
  AUTHORS   Xiao,Y., Qian,J., Hou,X., Zeng,L., Liu,X., Mei,G. and Liao,Y.
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1644)
  AUTHORS   Xiao,Y., Qian,J., Hou,X., Zeng,L., Liu,X., Mei,G. and Liao,Y.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-AUG-2022) Center of Economic Botany, Core Botanical
            Gardens, South China Botanical Garden, Chinese Academy of Sciences,
            Changxing, Guangzhou 510630, China
FEATURES             Location/Qualifiers
     source          1..1644
                     /organism="Ipomoea batatas"
                     /mol_type="mRNA"
```

Grants Awarded for International Collaboration

Funding Agency	Project Title	Collaborators	Amount
IUBS	Open Biodiversity and Health Big Data	Multiple countries	Euro 10,000
CAS	Global Genomics Data Sharing	USA	RMB 800,000
ANSO	Precision warning method for high-risk variants of emerging infectious diseases	Brazil, France, Pakistan	RMB 1,300,000
ANSO	Whole genome sequencing and miRNA biomarkers for an enhanced understanding of mechanism of tuberculosis infection in cynomolgus macaques (Macaca fascicularis): A translational knowledge to clinical study	Thailand, USA	US$ 150,000

CONTENTS

1 Updates

2 Development

3 Future directions

Future Directions

- Progress of CNCB infrastructure
- Collaboration with DDBJ and KOBIC
- Global Core Biodata Resource
- Partnership with INSDC
- Talent recruitment

Acknowledgement

Funding

- National Programs for High Technology Research and Development
- National Key Research Program of China
- Key Program of the Chinese Academy of Sciences
- Strategic Priority Research Program of Chinese Academy of Sciences
- IUBS
- ANSO

NGDC Members https://ngdc.cncb.ac.cn/people