WORKSHOP ON BIOINFORMATICS APPLICATIONS IN GENOMICS SURVEILLANCE OF BACTERIAL ANTIMICROBIAL RESISTANCE

Biological Databases and Resources

22/02/2023

Stephane Fadanka, Diapa Nana Yanick

Mboalab Biotech | Beneficial Bio

Stephane Fadanka

Executive Director, Mboalab Biotech
Managing Director (Cameroon), Beneficial Bio Ltd

- SynBio Africa Emerging Leader,
- OpenScience Ambassador (OLS-Cohort5),
- Mentor (Outreachy, OLS).

Stephane@beneficial.bio
@StephaneFadanka
697465154 /679475748

Diapa Nana Yanick

Director, Special projects and Innovation Ecosystem Mboalab Biotech,

- Sales Manager, Beneficial Bio Ltd

yanick@beneficial.bio
diapayanick2016@gmail.com

Biological Databases and Resources

Database or databank ?

- From now on, the term « database » (db) is usually preferred

What is a database?

- a database is defined as an organized collection of data or information that
- is electronically stored and accessible from a computer system.
- The organized nature of the database makes it easy to access, manage, periodically update, and rapidly search the required data/information from a suitable computer system.

Biological Databases and Resources

A collection of...

❑ structured
❑ searchable (index) -> table of contents
❑ updated periodically (release) -> new edition
❑ cross-referenced (hyperlinks) -> links with other db

…data

Includes also associated tools (software) necessary for db access, db updating, db information insertion, db information deletion….

Data storage management: flat files, relational databases…

Biological Databases and Resources

Among various types of databases, the ones constituting the datasets relevant to biological sciences such as molecular biology and bioinformatics are called **biological databases.** In the current scenario.

Importance:

- Rapidly advancing molecular biology, proteomics, and low-cost high-throughput genome sequencing technologies, huge amounts of biological information such as raw sequencing datasets, proteomes, etc. are being generated at a very rapid rate. **Thus, the storage and handling of this staggering information are the major challenges of the current genomics era.**

- Biological databases enables the scientists **to access and retrieve the biologically relevant data** including the raw data, genome sequences, analyzed datasets, and annotations in easily manageable/organized formats.

- Biological databases allow **data indexing as well as help remove data redundancy.** At present, biological databases have become the central component of bioinformatics. Through the various data mining tools, all biological information can be easily accessed; thus saving time, resources, and efforts.

Components of biological database

Similar to other databases, a biological database also has certain basic components:

a. **Entity** - An entity refers to the thing we want to store in a database. Eg. DNA sequences, Genes, Bibliographic references, etc.

b. **Fields** - The properties of an entity are called fields. Eg. Gene name, gene sequence, Mutation (if any), etc.

c. **Records** - A record typically refers to a combination of all the fields for a given entity. For eg. Record for gene BRCA1 in GenBank

d. **Identifier** - The unique name which identifies a record.

In the case of a simple database, a single file contains multiple records. Among these records, each one can have the same set of information (fields) along with a unique identifier.

Types of biological databases

Based on their content, the biological databases can be classified into the following types:

- **Primary databases**

Primary databases, also known as the archival databases, basically contain experimentally derived datasets such as nucleotide and protein sequences as well as the structural information of macromolecules. This basic information can be accompanied by functional annotation, bibliographies, and links to other databases. The data to the primary database is directly submitted by researchers. Once submitted, the data is assigned an accession number, which is permanent and becomes a part of the scientific record.

The followings are examples of primary databases:

1. Primary nucleotide sequence databases - The European Nucleotide Archive (ENA), The National Center for Biotechnology Information GenBank (NCBI GenBank), and The DNA Data Bank of Japan (DDBJ), etc.
2. Microarray/Functional genomics databases - Gene Expression Omnibus (GEO) and Array Express Archives etc.
3. Protein sequences and structure databases - Swiss-Prot and Protein Information Resource (PIR) for protein sequences, Protein Databank (PDB) for protein structure.

Types of biological databases

- **Secondary databases**

Secondary databases store the information derived from the analysis of primary datasets. Secondary databases contain highly curated information derived from complex computational as well as manual analysis of primary resources and scientific literature. These databases often store information about conserved domain structure/sequences, signal sequences, and active site residues.

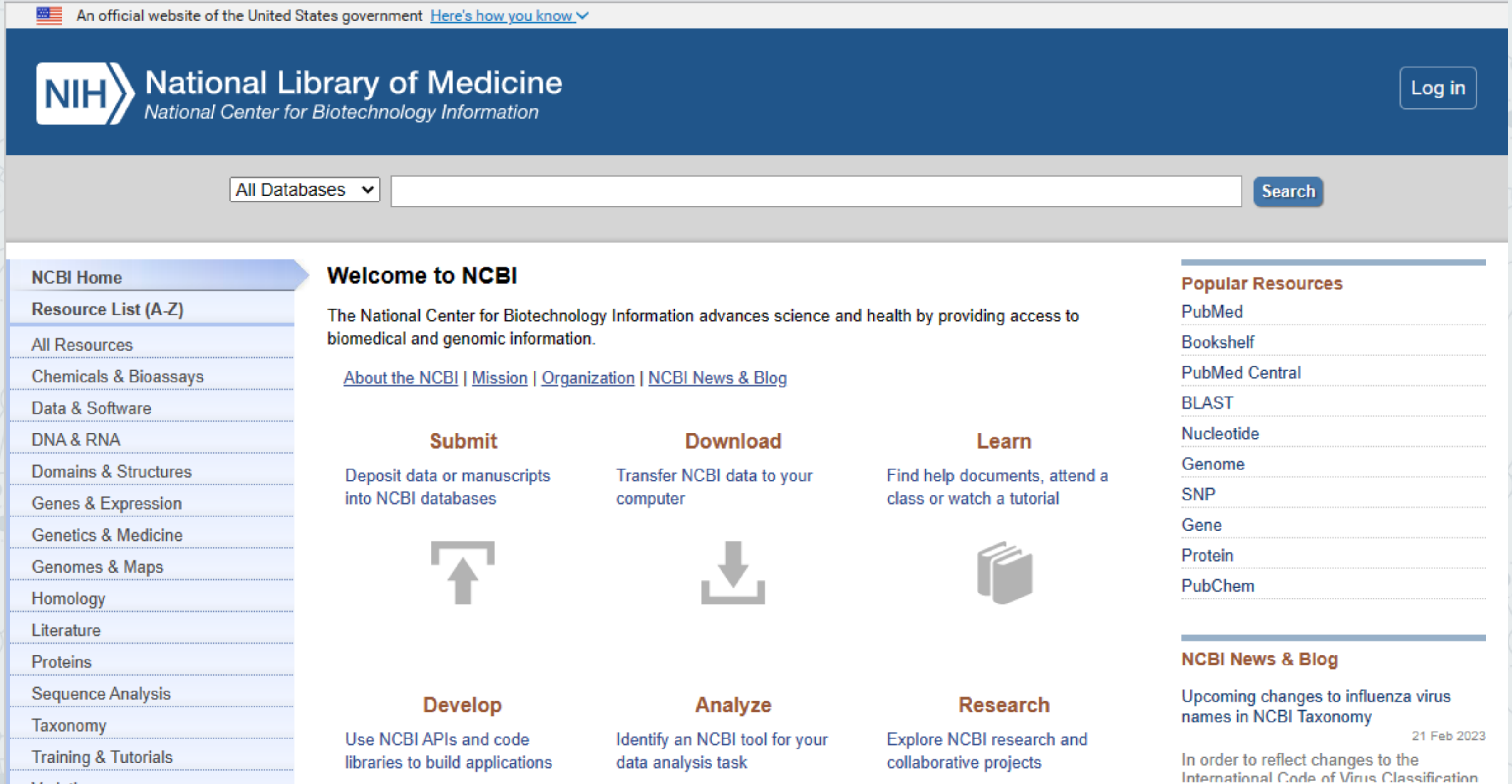
1. Protein families, domains and structure databases - InterPro, PROSITE, SCOP, CATH and NCBI Conserved Domain Database (CDD)
2. Protein sequences and functional information databases - UniProt Knowledgebase (UniProtKB)
3. Nucleotide (Genes/Genomes) sequence and annotation databases - NCBI UniGene, The European Bioinformatics Institute (EBI) Genomes (EBI Genomes), and Ensembl, etc.

- **Specialized databases**

These databases cater to the needs of specific research interests. Eg. Ribosomal Project Database (RDP), HIV sequence database, The Saccharomyces Genome Database (SGD), Mouse Genome Database (MGD), and Antibiotic Resistance Genes Database (ARDB), etc.

Important biological databases

A. Nucleotide sequence databases a. The National Center for Biotechnology Information GenBank (NCBI GenBank) NCBI is a part of the National Library of Medicine (NLM) under the U.S. National Institute of Health (NIH).



An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

All Databases | Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

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

NCBI News & Blog

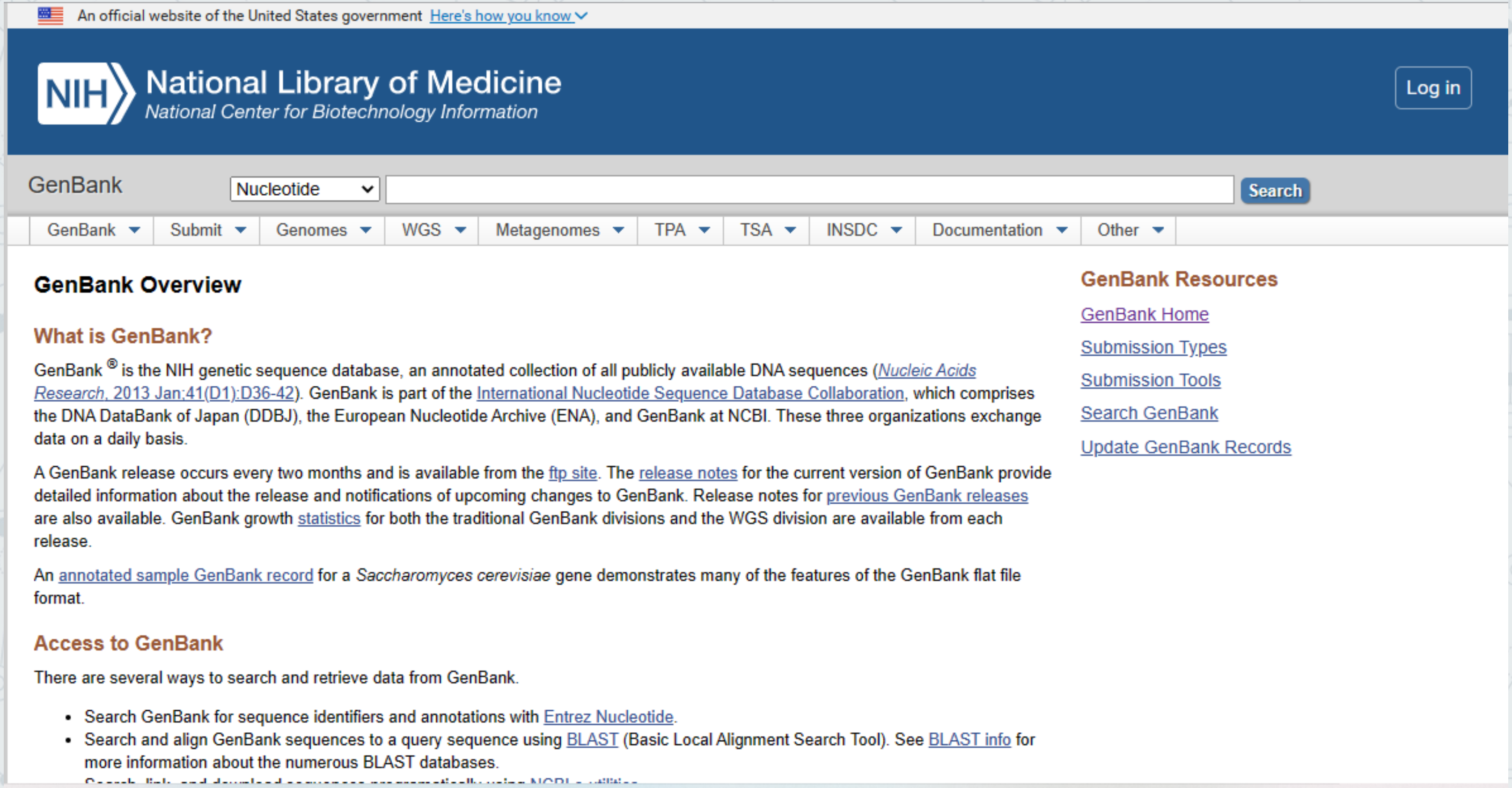
Upcoming changes to influenza virus names in NCBI Taxonomy

21 Feb 2023

In order to reflect changes to the International Code of Virus Classification

Important biological databases

GenBank is a publically available collection of nucleotide sequences, their protein sequences along with annotations



An official website of the United States government Here's how you know
NIH
National Library of Medicine
National Center for Biotechnology Information
Log in
GenBank
Nucleotide
Search
GenBank
Submit
Genomes
WGS
Metagenomes
TPA
TSA
INSDC
Documentation
Other

GenBank Overview

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research, 2013 Jan;41(D1):D36-42*). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample GenBank record for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with Entrez Nucleotide.
- Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). See BLAST info for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using NCBI e-utilities.

GenBank Resources

- GenBank Home
- Submission Types
- Submission Tools
- Search GenBank
- Update GenBank Records

Important biological databases

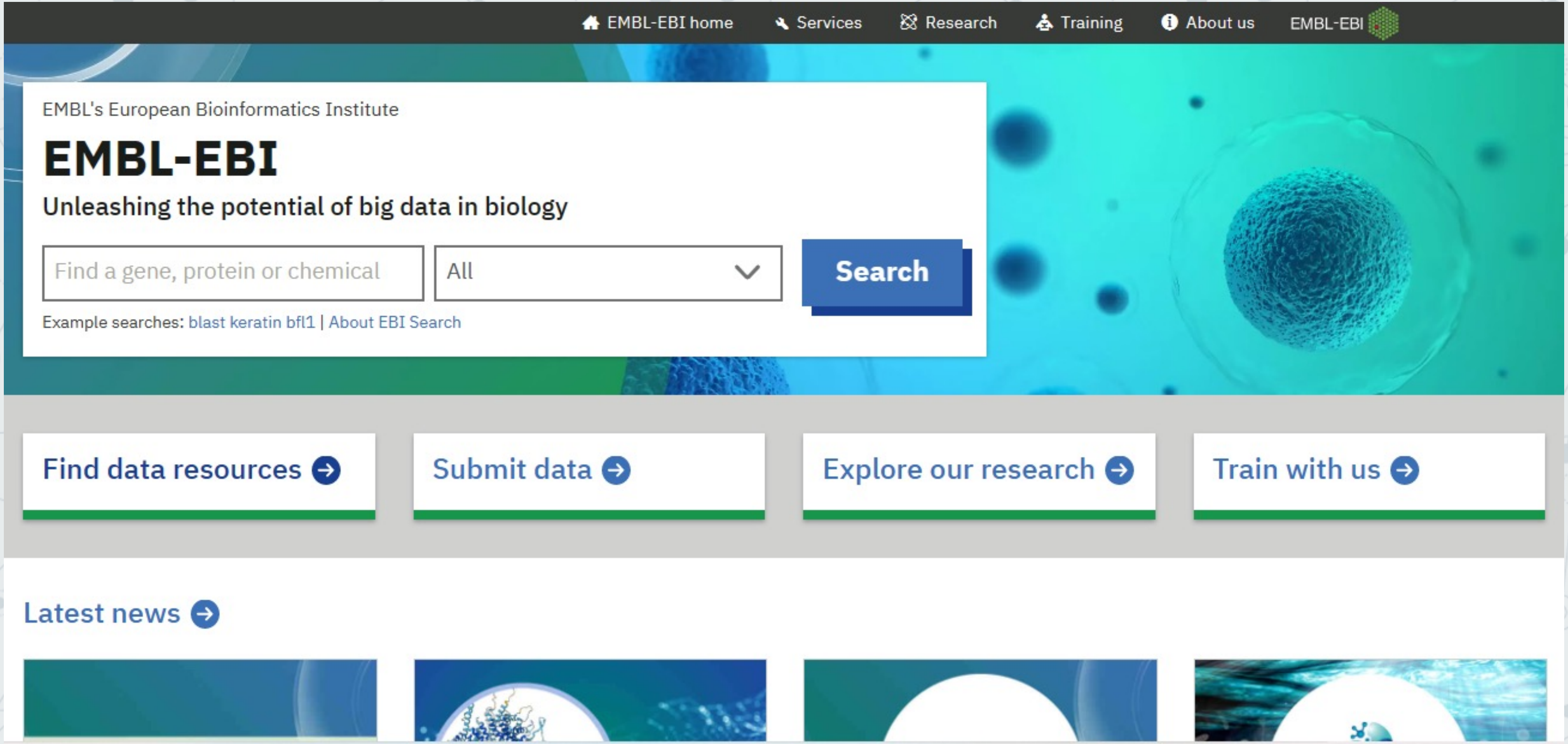
GenBank is a publically available collection of nucleotide sequences, their protein sequences along with annotations

Some of the unique features of Genbank are:

- ❑ Being a free public repository, any researcher can submit the sequences to GenBank without incurring any financial cost.
- ❑ For the same gene or genome, multiple sequences of varying quality are available in GenBank. Essentially, anything submitted to GenBank is stored.
- ❑ To represent the various modification done by the author, a sequence can have several versions.
- ❑ Once a record is submitted to GenBank, it is assigned to a specific division based on the source taxonomy or sequencing strategy used to obtain the data. There are 12 taxonomic divisions and 8 functional divisions.
- ❑ Each GenBank record containing a sequence is assigned a unique identifier called an accession number. The accession number is permanent, and it stays the same throughout the life of the record. Only changes may occur in the sequence version but not in the accession number. For eg. ACCESSION AF000001, VERSION AF000001.5

Important biological databases

B. EMBL: European Molecular Biology Laboratory: Similar to GenBank, the EMBL database (http://www.ebi.ac.uk/embl/index.html) maintained at European Bioinformatics Institute (EBI) is part of the European Nucleotide Archive (ENA) aimed at constructing a comprehensive catalog of the world's nucleotide sequencing information.



EMBL's European Bioinformatics Institute

EMBL-EBI

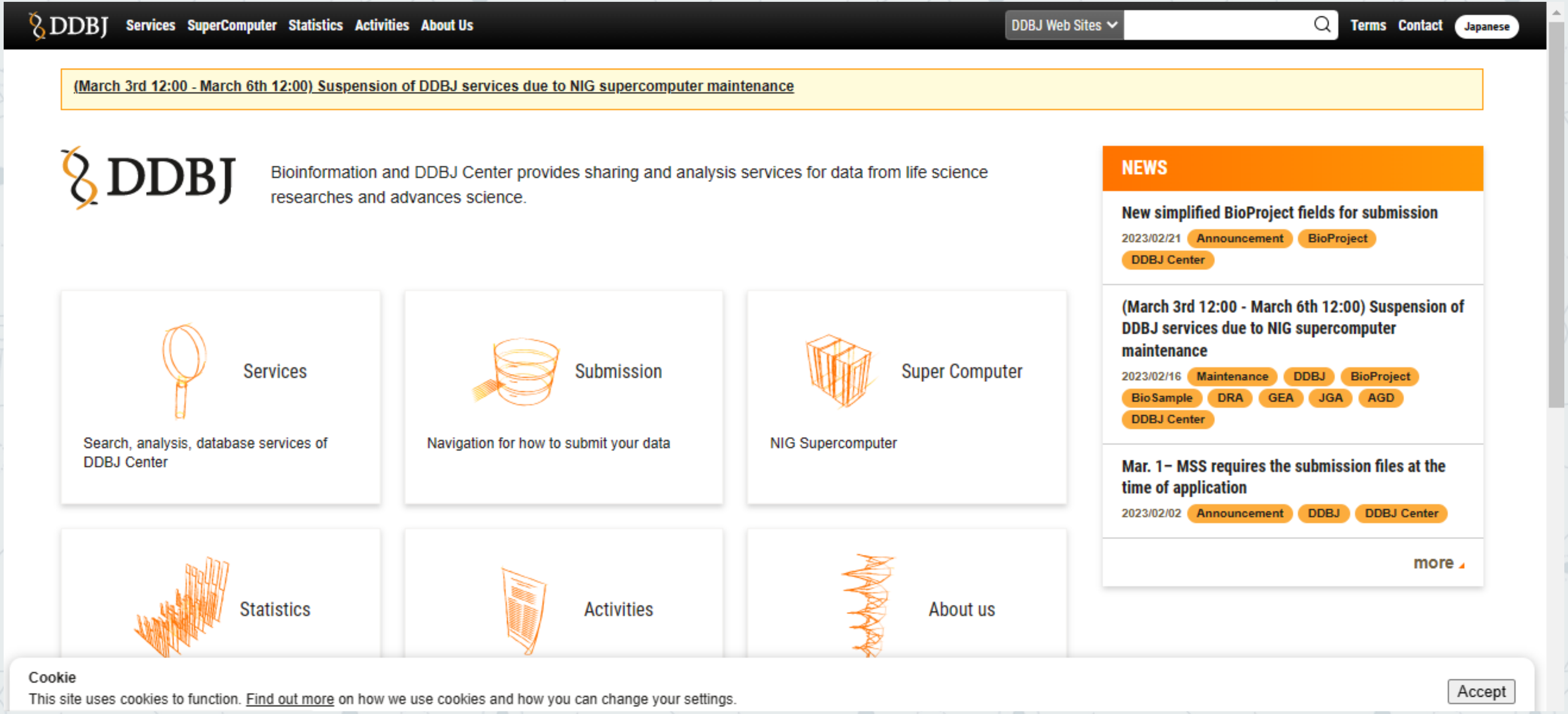
Unleashing the potential of big data in biology

Find a gene, protein or chemical | All

Example searches: blast keratin bfl1 | About EBI Search

Important biological databases

C. DNA Data Bank of Japan (DDBJ) DDBJ (https://www.ddbj.nig.ac.jp/) is the only nucleotide sequence database located in Asia. It was established in 1986 by the Center for Information Biology (CIB) under the National Institute of Genetics (NIG) of Japan in collaboration with NCBI in the USA and EBI in Europe. A



(March 3rd 12:00 - March 6th 12:00) Suspension of DDBJ services due to NIG supercomputer maintenance

DDBJ

Bioinformation and DDBJ Center provides sharing and analysis services for data from life science researches and advances science.

Services

Search, analysis, database services of DDBJ Center

Submission

Navigation for how to submit your data

Super Computer

NIG Supercomputer

Statistics

Activities

About us

NEWS

New simplified BioProject fields for submission
2023/02/21 Announcement BioProject DDBJ Center

(March 3rd 12:00 - March 6th 12:00) Suspension of DDBJ services due to NIG supercomputer maintenance
2023/02/16 Maintenance DDBJ BioProject BioSample DRA GEA JGA AGD DDBJ Center

Mar. 1– MSS requires the submission files at the time of application
2023/02/02 Announcement DDBJ DDBJ Center

more

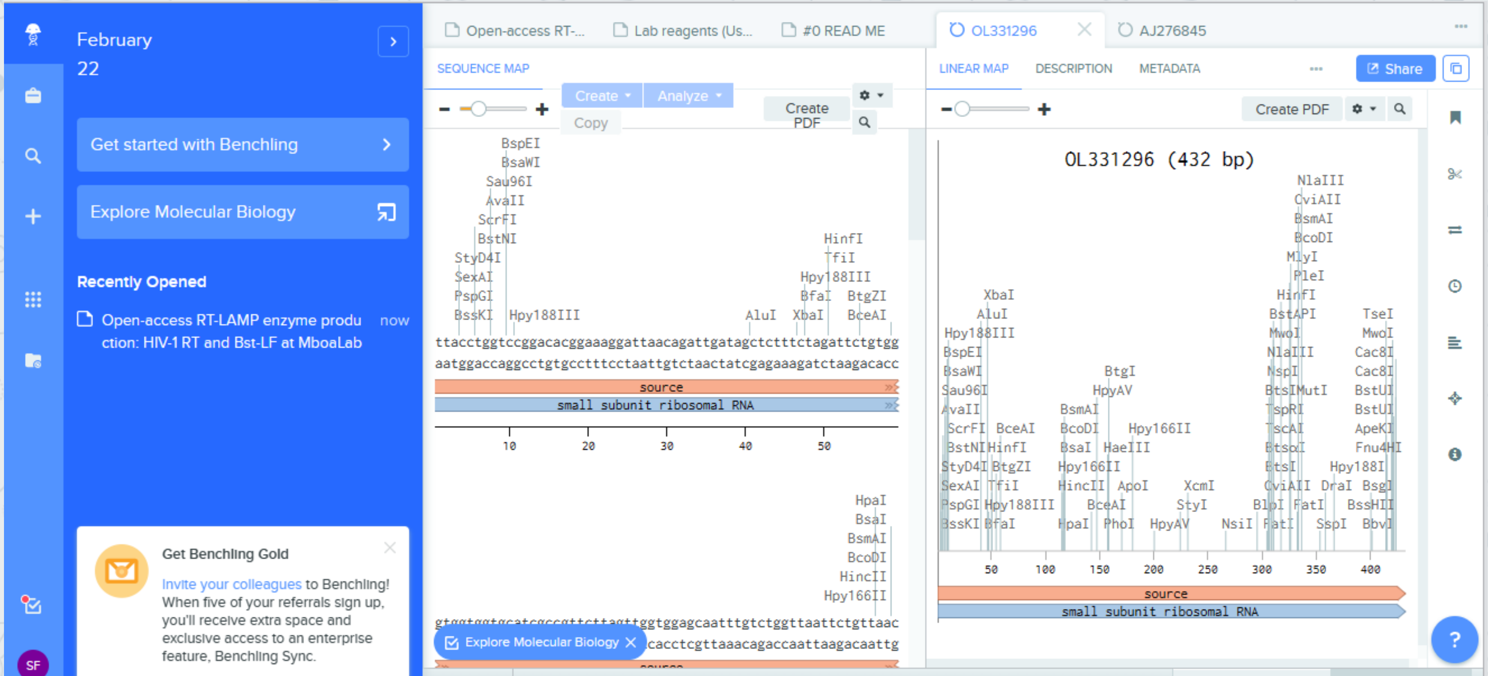
Cookie

This site uses cookies to function. Find out more on how we use cookies and how you can change your settings.

Bonus Ressource

Benchling

Benchling is a tool for virtually documenting your lab work. In addition to providing tools for writing and sharing protocols, it also has the functionalities for analyzing DNA and protein files, which is useful for *in silico* design of plasmids and oligos. Start by visiting Benchling and making a free account.



Open-access RT-LAMP enzyme production: HIV-1 RT and Bst-LF at MboaLab now

```
ttacctggtccggacacggaaaggattaacagattgatagctctttctagattctgtgg
aatggaccaggcctgtgcctttcctaattgtctaactatcgagaaagatctaagacacc
```

```
gtggtggtgcatcgccgttcttagttggtggagcaatttgtctggttaattctgttaac
Explore Molecular Biology cacctcgttaaacagaccaattaagacaattg
```

Get Benchling Gold

Invite your colleagues to Benchling! When five of your referrals sign up, you'll receive extra space and exclusive access to an enterprise feature, Benchling Sync.

Thanks!

MBOA LAB

BENEFICIAL.BIO

Stephane Fadanka

- ★ Stephanefadanka@gmail.com
- ★ https://github.com/Fadanka
- ★ @StephaneFadanka

Diapa Nana Yanick

- ★ yanick@beneficial.bio
- ★ diapayanick2016@gmail.com