

INSDC の塩基配列データベースを知る・使う・その舞台裏

児玉 悠一

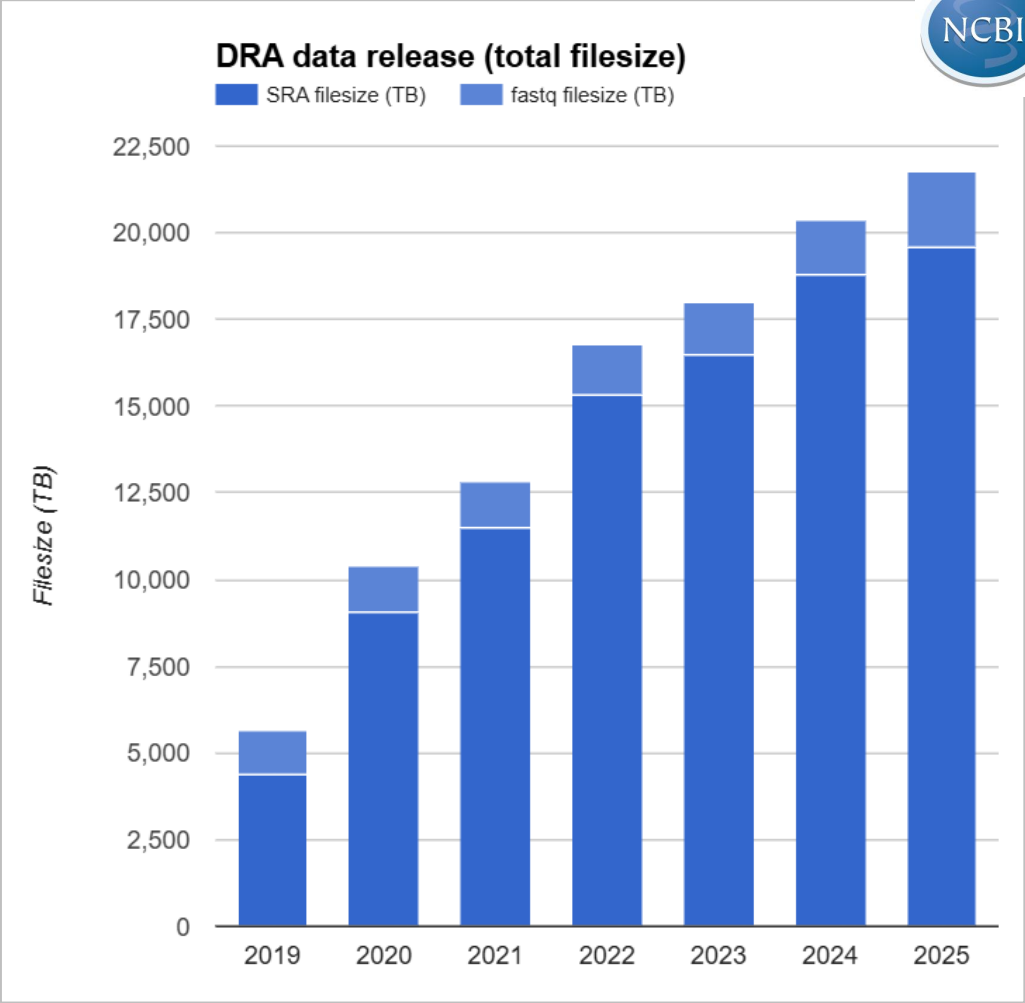
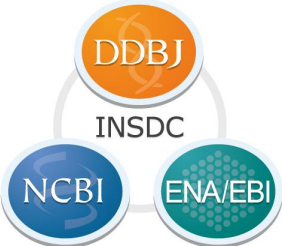
国立遺伝学研究所 バイオデータ研究拠点 (BSI)

2. 国際塩基配列データベースの舞台裏

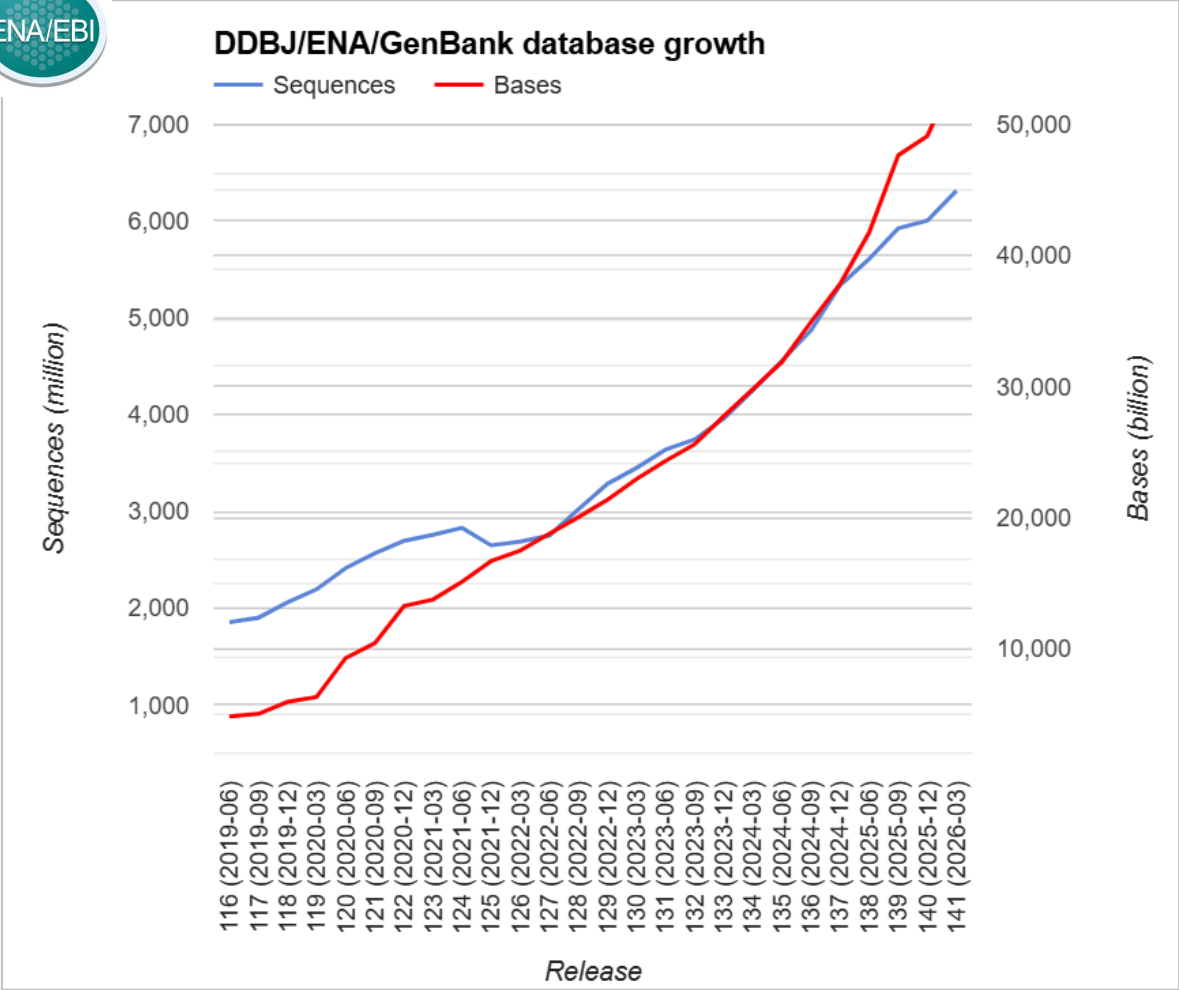
- 増大するデータ・透明化・社会的責任

増大するデータ

増え続けるデータの量と数

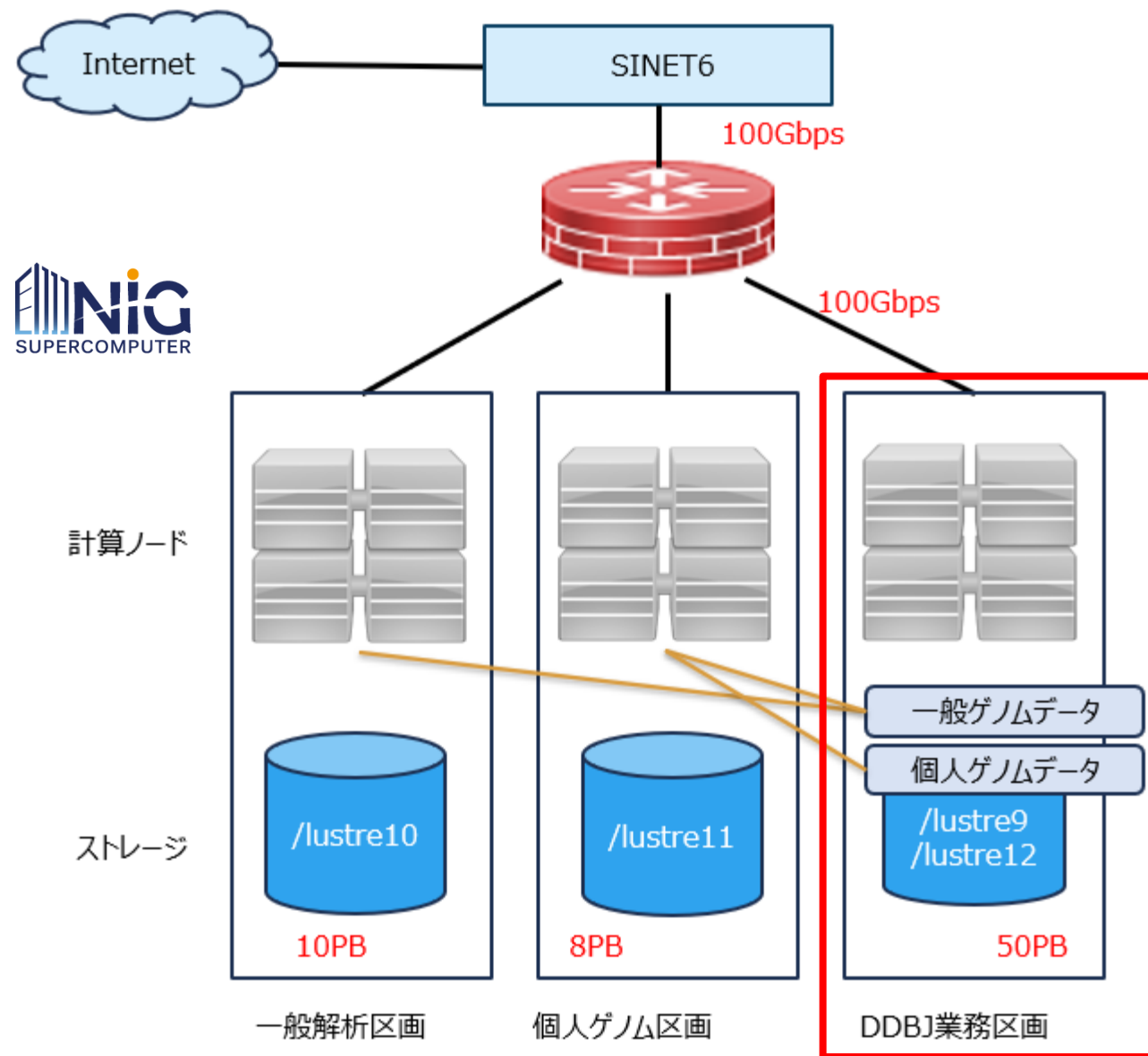


4京6,500兆 塩基, 22ペタバイト (量)

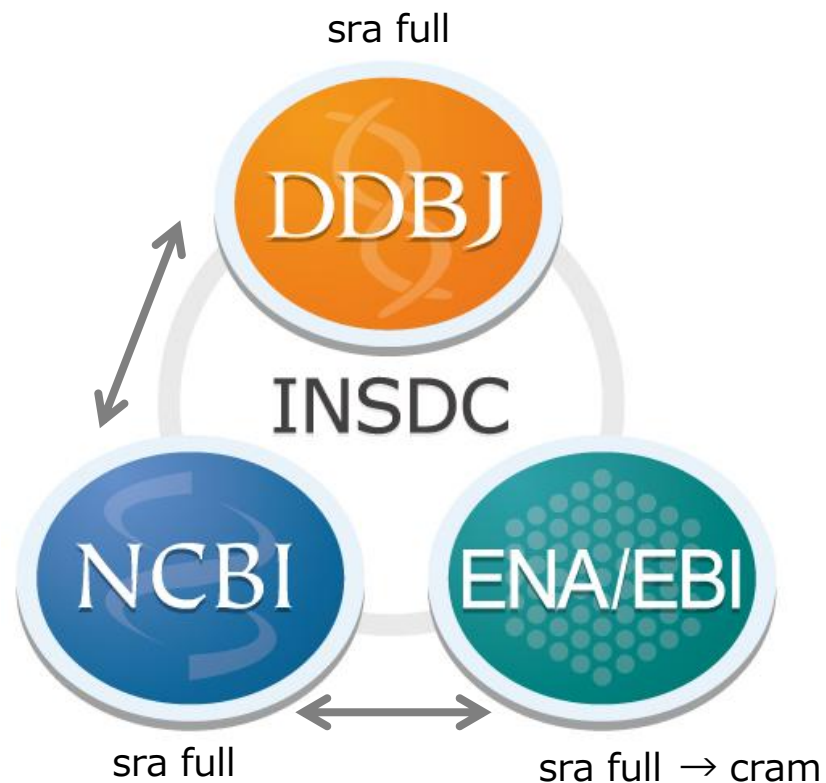


67億 塩基配列, 10テラバイト (数)

ディスク容量



量への対応①: Quality value (QV) の圧縮



sra, full-scale QVs → sra lite, per read QVs を交換

Full-scale QVs を利用したい場合は、受付け元から公開されている submitted files（登録された元ファイル）にアクセスする

INSDC

DRR1080205.sra.lite

DRR1080206.sra.lite

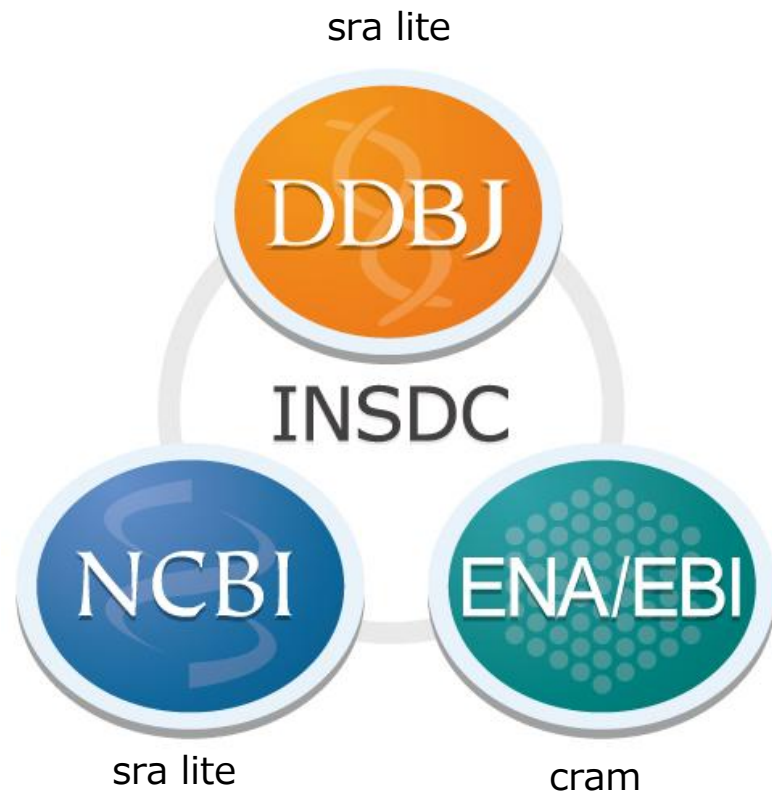
DRA

illumina_1.fq.gz, illumina_2.fq.gz

pacbio.fq.gz

注意: ディスク容量が許す限り full を提供する方針であり、すぐに lite に移行するわけではない

量への対応②: フェデレーション



メタデータは全て交換

sra lite は受付け元以外の INSDC メンバーにミラーされているとは限らない

メタデータは提供されている sra lite をリンク

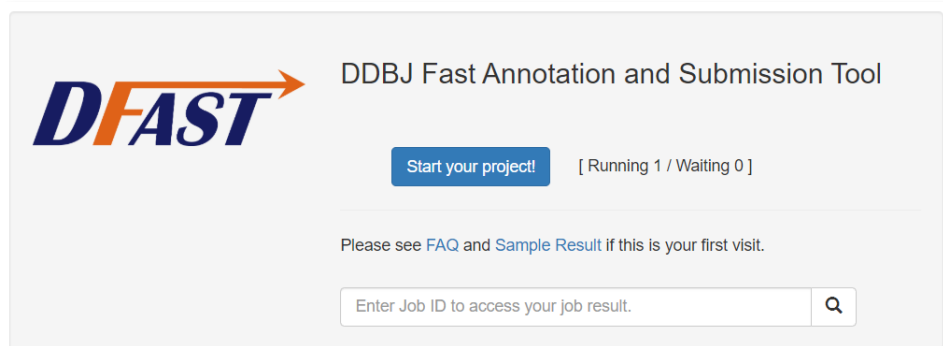
DRA

SRR999999999 Run metadata → SRR999999999.sra.lite at NCBI SRA

注意

- ・ ディスク容量が許す限り全 sra ファイルを交換する方針であり、すぐにフェデレーションに移行するわけではない
- ・ アセンブリ配列は INSDC メンバー (DDBJ/ENA/GenBank) で全て共有

数への対応 DDBJ: 登録の自動化



- DFAST 原核生物ゲノムアノテーションパイプライン
- MSS フォームで DFAST job ID を指定すると、出力されたアノテーション・配列ファイルが取り込まれる
- DDBJ の原核生物ゲノム登録の90%以上は DFAST 経由

数への対応: GenBank のアノテーションパイプライン

★ What type of organism or material was sequenced?

☒ Prokaryote ? ☐ Eukaryote ? ☐ Virus ?

★ What do your sequences contain?

☐ small subunit rRNA only (16S rRNA) (example records)
☐ large subunit rRNA only (23S rRNA) (example records)
☐ intergenic spacer (16S-23S rRNA IGS) (example records)
☐ any other feature (coding regions, regulatory regions, etc.)

★ What type of organism or material was sequenced?

☐ Prokaryote ? ☒ Eukaryote ? ☐ Virus ?

★ What do your sequences contain?

☐ nuclear ribosomal RNA or rRNA-ITS (example records)
☐ mitochondrial or chloroplast rRNA (example records)
☐ mitochondrial COX1 from metazoa only (example records)
☐ any other feature (coding regions, organelle genomes, regulatory regions, etc.)

★ What type of organism or material was sequenced?

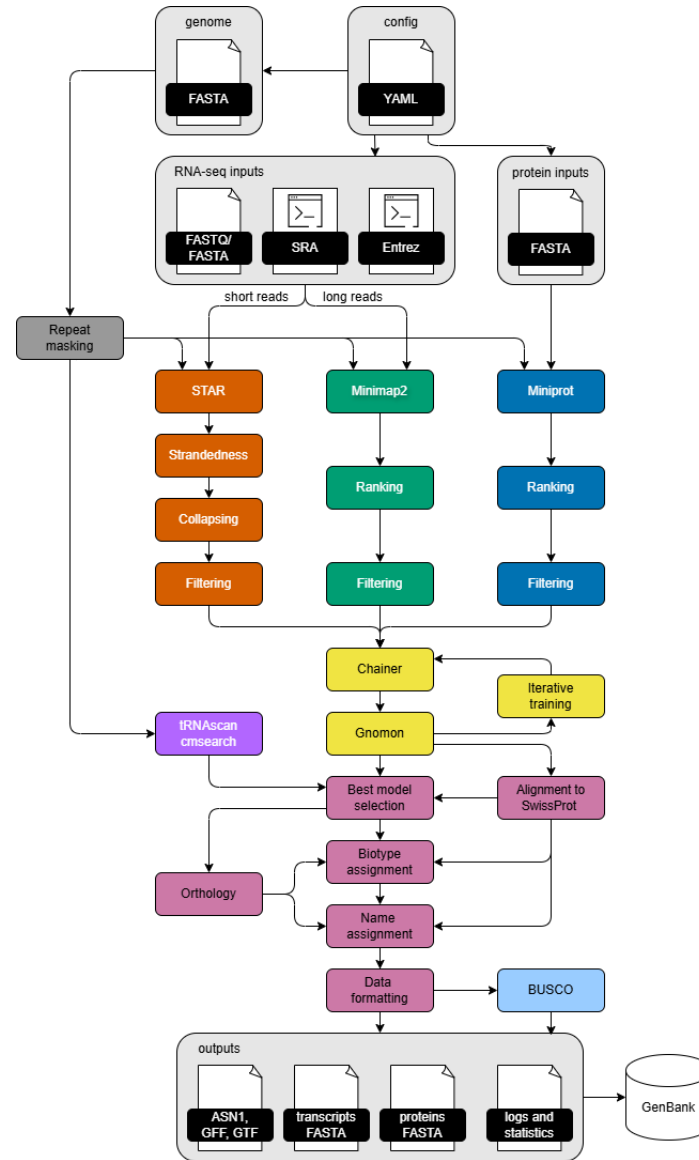
☐ Prokaryote ? ☐ Eukaryote ? ☒ Virus ?

★ What type of virus did you sequence?

☐ SARS-CoV-2 (example records)
☐ Influenza virus (example records)
☐ Norovirus (example records)
☐ Dengue virus (example records)
☐ any other virus (examples: HIV-1, Rotavirus, Herpesvirus, phage, viroid, etc.)

- rRNA
- 後生動物ミトコンドリア COX1 遺伝子
- 病原ウイルスゲノム (SARS-CoV-2、インフルエンザなど)
- 原核生物ゲノム (PGAP)

NCBI Eukaryotic Genome Annotation Pipeline (EGAP)



GenBank で EGAP を使用した真核生物ゲノム登録が増えている

限られた人員と予算

データベースセンターの規模



予算: 181億
人員: 600人

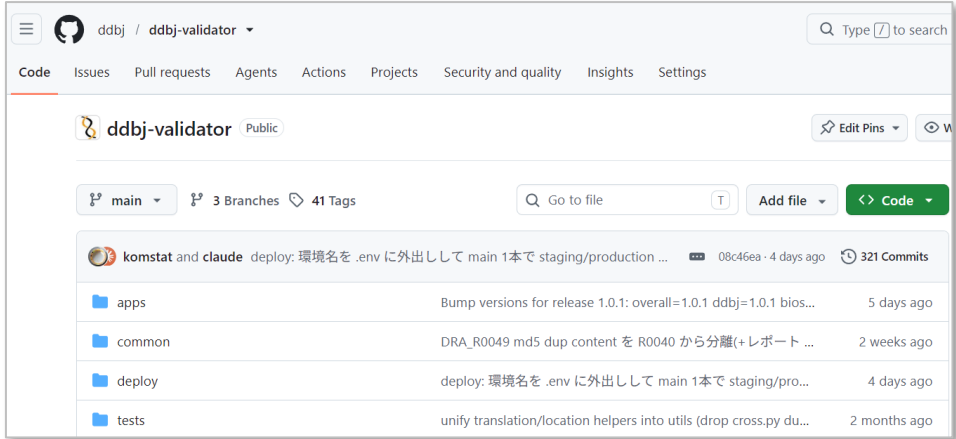


予算: 95億
人員: 780人



予算: 9.5億
人員: 44人

AI を使って統合 validator を自前開発



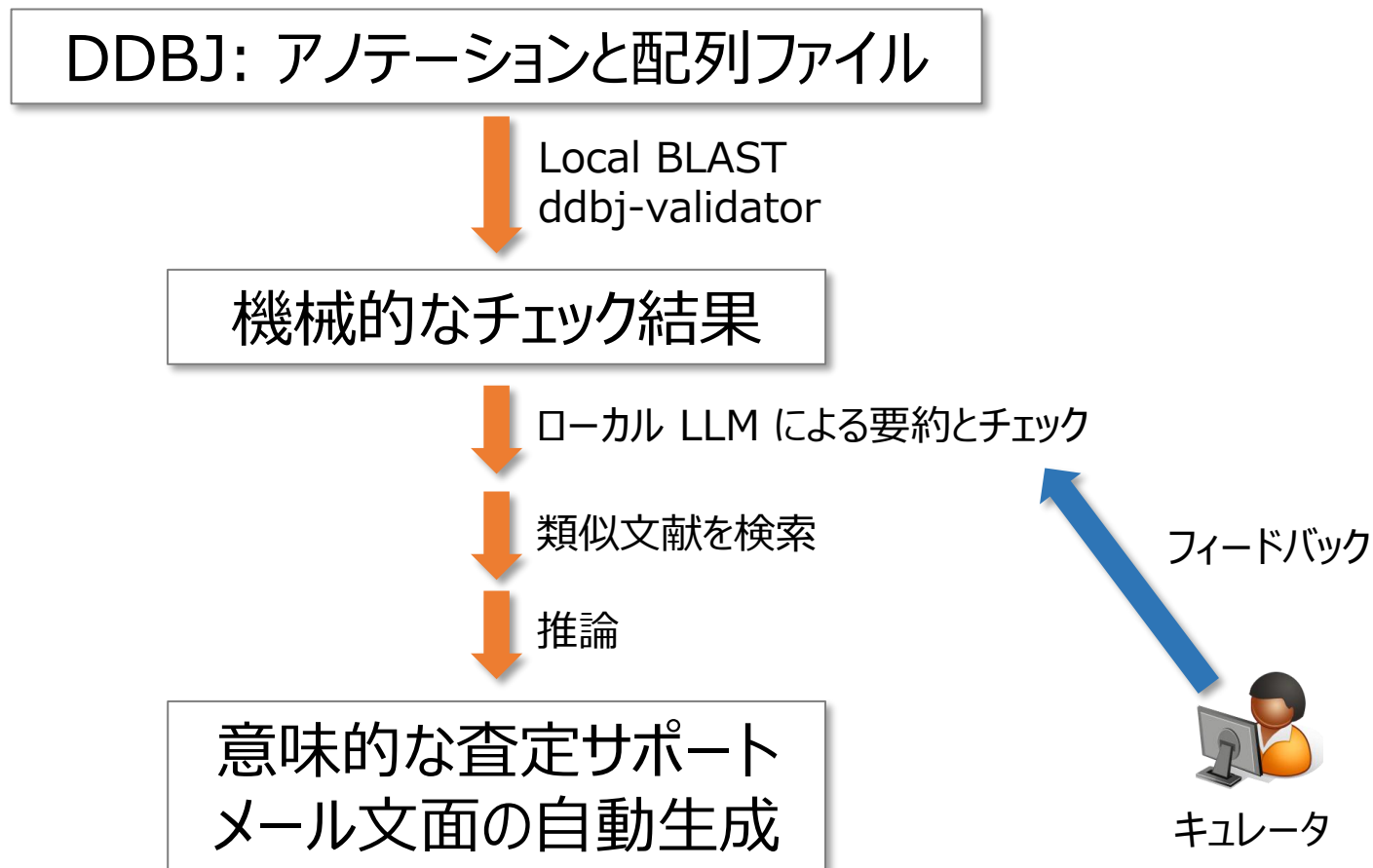
<https://github.com/ddbj/ddbj-validator>

データベース	ルール数
BioProject	29
BioSample	90
DRA	45
DDBJ	225
GEA	118
MetaboBank	43

Rule ID	Alternate IDs	Level	Min Spec	Autofix/cleanup	Submission	External DB/API	Auth	Message
ANN0010		FATAL						No annotation or sequence files found.
ANN0020	FMT0002	FATAL						Invalid file pair for '{basename}': {annCount} annotation file(s) and {seqCount} sequence file(s).
ANN0035		WARNING		auto-cleanup				[Auto-cleanup] Line endings will be automatically converted to LF.
ANN0040	FMT0004	ERROR						Non-ASCII characters detected.
ANN0050	FMT0005	ERROR						Line exceeds 10,000 characters.
AXS0060	JP0009	ERROR						Entry count mismatch: annotation ({N1}) and sequence ({N2}).
AXS0070	JP0011	ERROR						Entry name mismatch: annotation {ENTRY NAME1} and sequence {#ENTRY NAME2}.
SEQ0080	JP0013	ERROR						Missing FASTA definition line (>Entry name).

- AI を使って validator を自前開発
- BioProject/BioSample/DRA/DDBJ/GEA/MetaboBank 統合 validator（実装ルール 計550）
- python オープンソース
- 全ルールリストを公開
- D-way BioSample 登録システムで運用開始
- アノテーションファイルと配列ファイルの事前チェックに使える（UME よりも高機能）

大規模言語モデル（LLM）による査定サポート



データベースセンターの規模



予算: 181億
人員: 600人



予算: 95億
人員: 780人



予算: 9.5億
人員: 44人



2026年4月 組織統合
バイオデータ研究拠点 (BSI) へ

透明化

各国で立ち上がる塩基配列データベース



NGDC/CNCB/CAS
中国 北京



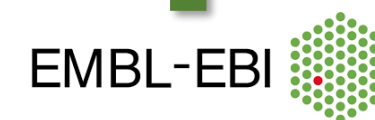
KOBIC/KRIBB
韓国 大田



INNA/BRIN
インドネシア ボゴール



IBDC/DBT
インド ニューデリー近郊



INSDC 加入を支援

PRJDB40828 (PRJCA007585 でも検索可能)

DDBJ Search Home > Entries > BioProject > PRJDB40828

Title	HS and BIN2 (PRJCA007585)
Name	Arabidopsis
Description	Identification of BIN2 and HS co-regulated genes.
Organism	Arabidopsis [?]
Publications	Heat Shock-Induced Accumulation of the Glycogen Synthase Kinase 3-Like Kinase BRASSINOSTEROID INSENSITIVE 2 Promotes Early Flowering but Reduces Thermotolerance in Arabidopsis. [?]
Grants	missing 32070293 / National Natural Science Foundation of China (NSFC) missing C2020205028 / Science and Technology Department of Hebei Province missing GCC2014002 / Education Department of Hebei Province
Organizations	Hebei Normal University
External Links	NGDC BioProject PRJCA007585 [?]
Project Type	Other



BioProject
BioSample
KRA

PRJCA007585

BioProject

Biological Project Library

Find a bioproject accession

e.g. PRJCA000028, PRJNA104308, human

EN

English

Accession

PRJCA007585

Title

HS and BIN2

Relevance

Model organism

Data types

Whole genome sequencing

Organisms

[Arabidopsis](#)

Description

Identification of BIN2 and HS co-regulated genes.

Sample scope

Multisite

Release date

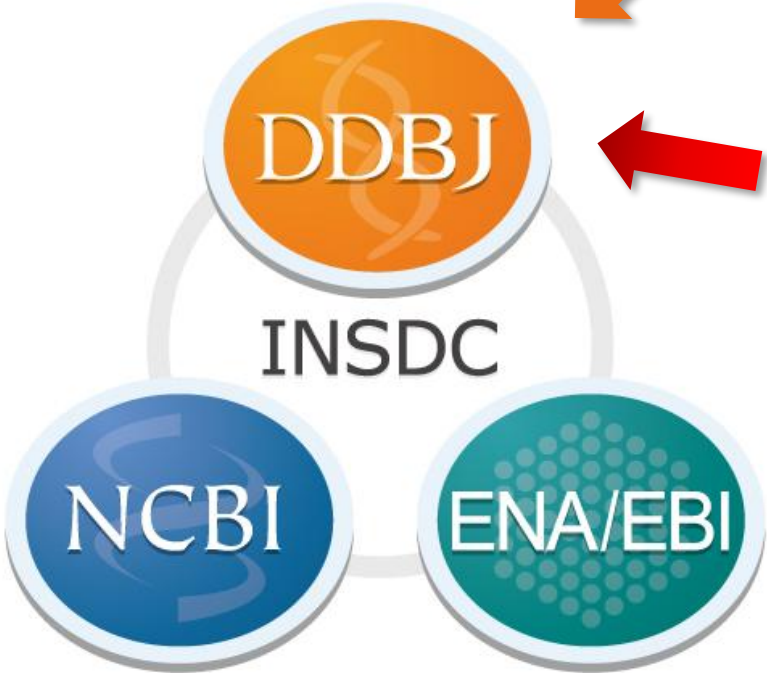
2025-12-10

Publication

Published ID	Article title	Journal name	DOI	Year
20154235	Heat Shock-Induced Accumulation of the Glycogen Synthase Kinase 3-Like Kinase BRASSINOSTEROID INSENSITIVE 2 Promotes Early Flowering but Reduces Thermotolerance in Arabidopsis	Frontiers in Plant Science	10.3389/fpls.2022.830602	2022



BioProject
BioSample
GSA



PRJDB40828

HS and BIN2 (PRJCA007585) Accession: PRJDB40828 ID: 1459720

Identification of BIN2 and HS co-regulated genes.

Accession	PRJDB40828												
Scope	Multisite												
Organism	Arabidopsis [Taxonomy ID: 3701] Eukaryota, Viridiplantae, Streptophyta, Embryophyta, Tracheophyta, Spermatophyta, Magnoliopsida, eudicotyledons, Gunnerales, Pentapetalae, rosids, malvids, Brassicales, Brassicaceae, Camellineae, Arabidopsis												
Publications	Ren H et al. "Heat Shock-Induced Accumulation of the Glycogen Synthase Kinase 3-Like Kinase BRASSINOSTEROID INSENSITIVE 2 Promotes Early Flowering but Reduces Thermotolerance in Arabidopsis", Front Plant Sci. 2022;13:830602.												
Grants	• "missing" (Grant ID C2020205028, Science and Technology Department of Hebei Province) • "missing" (Grant ID 32070293, National Natural Science Foundation of China (NSFC)) • "missing" (Grant ID GCC2014002, Education Department of Hebei Province)												
Submission	Registration date: 29-Apr-2026 Hebei Normal University												
Related Resources	• NGDC BioProject PRJCA007585												
Project Data:	<table><thead><tr><th>Resource Name</th><th>Number of Links</th></tr></thead><tbody><tr><td>SEQUENCE DATA</td><td></td></tr><tr><td>SRA Experiments</td><td>9</td></tr><tr><td>PUBLICATIONS</td><td></td></tr><tr><td>PubMed</td><td>1</td></tr><tr><td>PMC</td><td>1</td></tr></tbody></table>	Resource Name	Number of Links	SEQUENCE DATA		SRA Experiments	9	PUBLICATIONS		PubMed	1	PMC	1
Resource Name	Number of Links												
SEQUENCE DATA													
SRA Experiments	9												
PUBLICATIONS													
PubMed	1												
PMC	1												

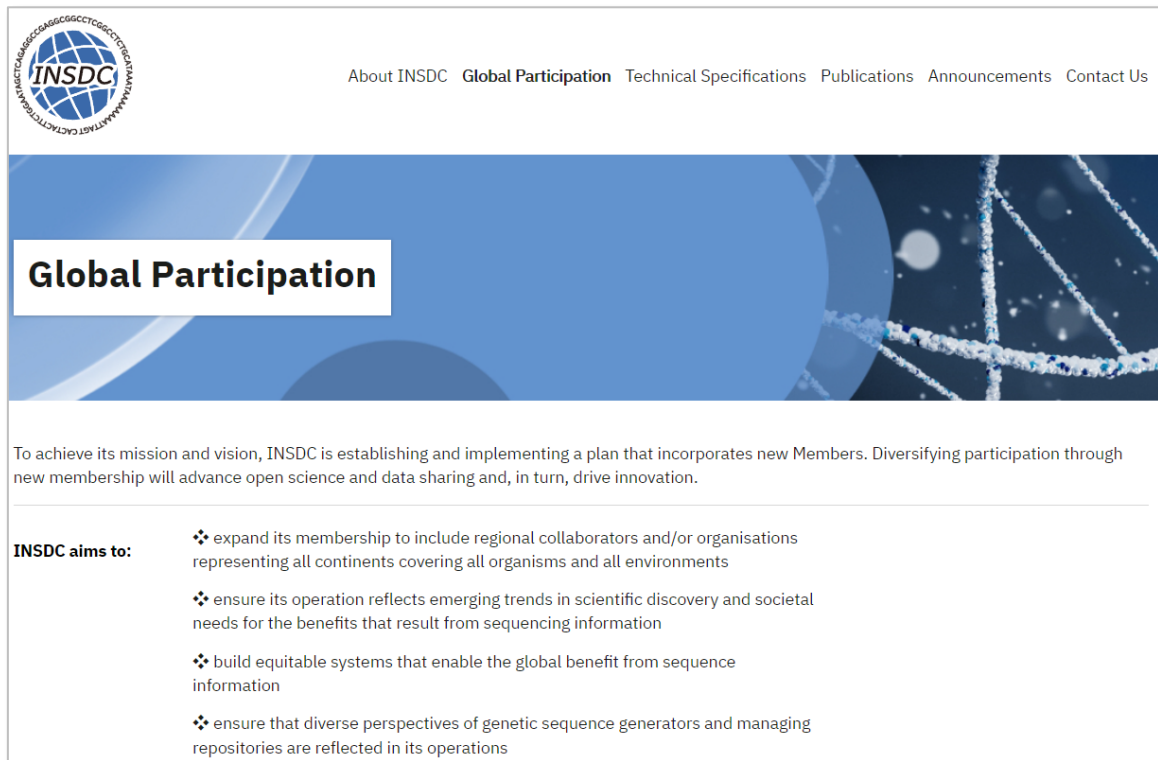
PRJDB40828

Project: PRJDB40828

Identification of BIN2 and HS co-regulated genes.

Organism:	Arabidopsis
Secondary Study Accession:	DRP018025
Study Title:	HS and BIN2 (PRJCA007585)
Center Name:	Hebei Normal University
Study Name:	Arabidopsis

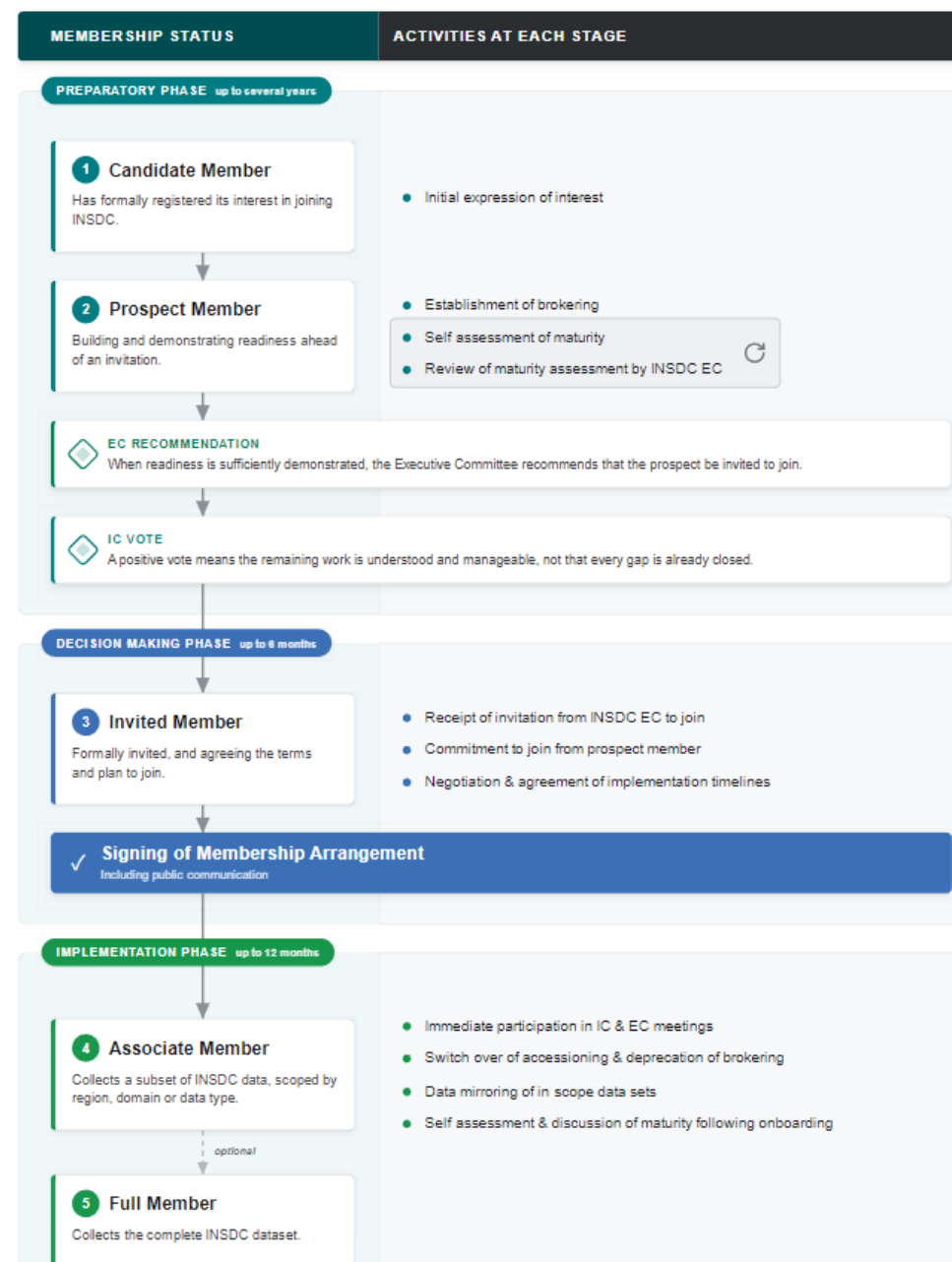
INSDC グローバル参画



The screenshot shows the INSDC Global Participation page. At the top is the INSDC logo and a navigation bar with links: About INSDC, Global Participation, Technical Specifications, Publications, Announcements, and Contact Us. Below the navigation bar is a large blue banner with the text "Global Participation". Underneath the banner, a paragraph states: "To achieve its mission and vision, INSDC is establishing and implementing a plan that incorporates new Members. Diversifying participation through new membership will advance open science and data sharing and, in turn, drive innovation." Below this paragraph, under the heading "INSDC aims to:", there are four bullet points:

- ❖ expand its membership to include regional collaborators and/or organisations representing all continents covering all organisms and all environments
- ❖ ensure its operation reflects emerging trends in scientific discovery and societal needs for the benefits that result from sequencing information
- ❖ build equitable systems that enable the global benefit from sequence information
- ❖ ensure that diverse perspectives of genetic sequence generators and managing repositories are reflected in its operations

- INSDC のミッションを明文化
- INSDC 加入までの道筋を提示
- Associate member (BioProject/BioSample/SRA)
- Full member (+ assembled sequences)



INSDC Minimal Specifications



About INSDC Global Participation Technical Specifications Publications Announcements Contact Us

INSDC Minimal Specifications

The INSDC minimal specifications define the core technical and data requirements that all INSDC databases must meet to ensure consistent, interoperable, and reliable sharing of nucleotide sequence data. They were formalised by the INSDC Implementation Committee in 2025, with the objective of reconciling historical standards and validation practices with modern data generation and use. Variations in how specifications had been implemented across member databases highlighted the need to clearly articulate a common set of requirements and the broader goals of the INSDC Global Participation. These specifications are intended to be living documents that will evolve over time in response to the needs of the global data community.

Category Name: Annotation

Definition

Description and location of the biological features in the cited sequence.

Minimum validation requirements

a. Syntactic validation

- Must conform to the [INSDC GFF3 specification](#) or [INSDC feature table](#), as appropriate for the data type (e.g., GFF3 for genome annotation on sequences in scope for an assembly object and assigned a GCA accession)

b. Annotation content validation

- Features and qualifiers must be limited by definitions and standards in the [INSDC Feature Table](#).
- Exceptions to feature validation rules exist for specific conditions, and are indicated with * below with further details described in table 2e
- Feature location:
 - Location descriptors (start/end) and strand must be provided for all features.
 - Feature spans must be within the sequence length*.

INSDC Availability

Shared

Data Format(s)

- GFF3 following INSDC syntactical specification, for agreed data types (e.g. genomes)
- Feature table following INSDC syntactical specification.

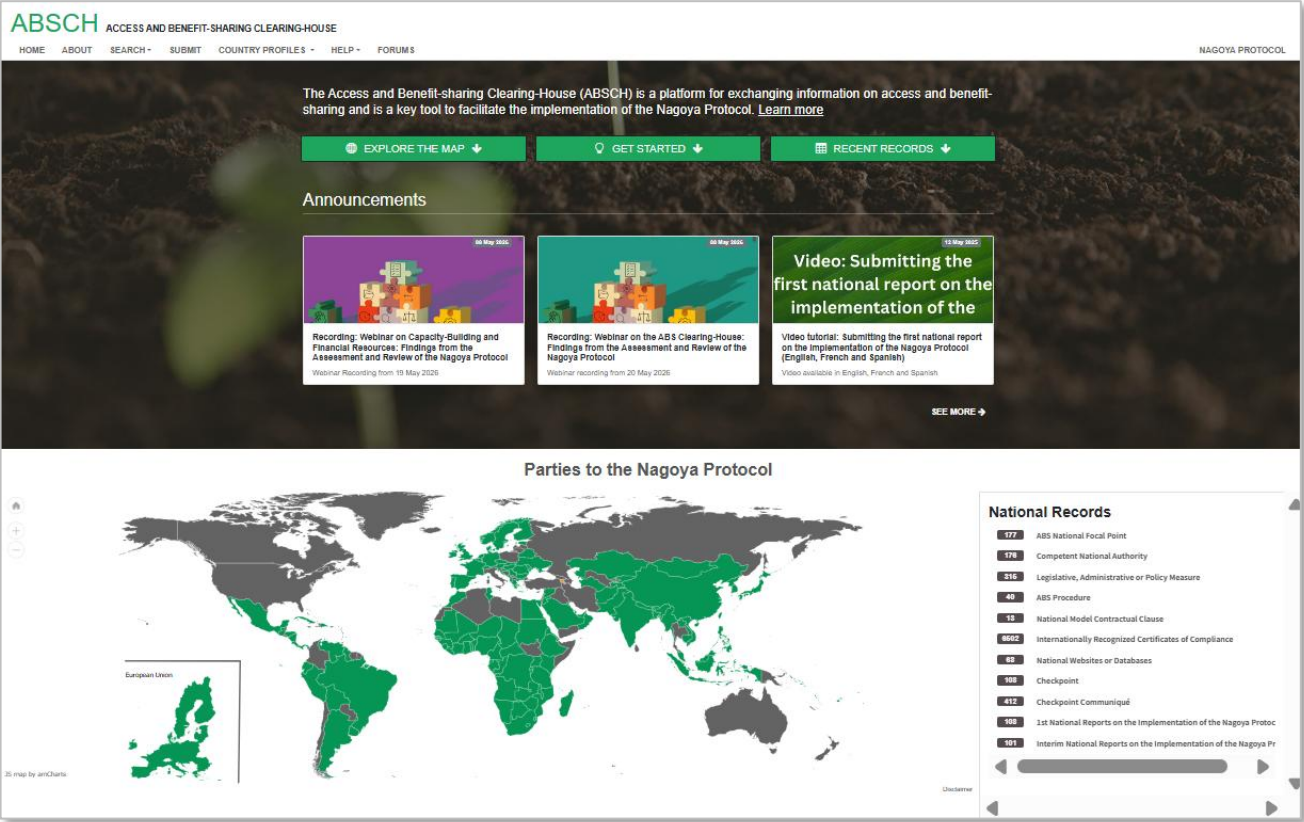
Tools

- syntactic GFF3 validation:
<http://www.sequenceontology.org/software/GAL.html> - Perl (unsupported)
<http://genometools.org/cgi-bin/gff3validator.cgi>
<https://github.com/NAL-i5K/GFF3toolkit> - Python
and the command line version: gt gff3validator which provides functionality to specify the allowable sequence ontology terms and to validate desired cross reference identifier formats.
- table2asn can read GFF3

INSDC 各カテゴリの最小要件を定義

- Project
- Sample
- Package-Checklist
- Analysis
- Experiment
- Raw Reads
- Compressed Reads
- Annotations
- Assembly
- Sequence

社会的責任



Wildlife & Biodiversity

Biopiracy 2.0: Will the proposed multilateral benefit sharing mechanism legitimise another great gene plunder?

Without strong safeguards, the multilateral mechanism for digital sequence information could deepen the divide between developing and developed countries

Ramakrishnan, N. (2024) Down To Earth, 18 Oct 2024.
<https://www.downtoearth.org.in/wildlife-biodiversity/biopiracy-20-will-the-proposed-multilateral-benefit-sharing-mechanism-legitimise-another-great-gene-plunder>

INSDC のような配列情報を自由に共有しているデータベースが
適正な ABS (Access and Benefit-Sharing, 遺伝資源へのアクセスと利益配分) の妨げになっているのではないか？

INSDC の対応①: サンプル採取地と採取日の必須化

2023年5月 BioSample でサンプル採取地 (geo_loc_name) と採取日 (collection_date) を必須化

2024年末 DDBJ においても必須化

2023年5月の BioSample での必須化前後

前1年間: 約70% のサンプルで geo_loc_name と collection_date が記載されていた

後1年間: 90%以上のサンプルで記載されるようになり、充足率が向上

DDBJ の対応②：同意チェックボックスの設置

Article | [Open access](#) | Published: 24 June 2026

How can biological databases support the new UN mechanism for benefit-sharing from digital sequence information?

[Débora S. Raposo](#), [Davide Faggionato](#), [Barbara Ebert](#), [Pablo Orozco](#), [Guy Cochrane](#), [Yiming Bao](#), [Masanori Arita](#), [Melania Muñoz-García](#), [Jens Freitag](#), [Tim Hirsch](#), [Charles E. Cook](#), [Rutger Vos](#), [Lorenz Reimer](#), [Sujeevan Ratnasingham](#), [Mathieu Rouard](#), [Sarah Brinkley](#), [Matthias Lange](#), [Uwe Scholz](#), [Stephan Weise](#), [Björn Usadel](#), [Robert D. Finn](#), [Chan-Ho Park](#), [Aylin S. Haas](#), [David Castle](#), [Colman O’Cathail](#) & [Amber H. Scholz](#)  [— Show fewer authors](#)

Scientific Data **13**, Article number: 971 (2026) | [Cite this article](#)

SUBMITTERGENERAL INFOSAMPLE PACKAGEATTRIBUTESOVERVIEW

SUBMITTER

Submitter 1

First name	Mishima
Last name	Hanako
Email	test09@nig.ac.jp

Submitter 2

First name	Ichiro
Last name	Fuji
Email	test10@nig.ac.jp

Organization

Submitting organization	National Institute of Genetics
Submitting organization URL	

GENERAL INFO

Data Release

Data Release	Hold
--------------	------

External Links 1

Link description	
URL	

Private comments to DDBJ staff

Comments	
----------	--

SAMPLE PACKAGE

Core Package	MiXs compliant
MiXs Sample	Eukaryotic Genomic Sequences (MIGS)
MiXs Sample	No environmental package

ATTRIBUTES

Attributes

TSV file	SSUB048035.txt
----------	----------------

☒ I confirm that, to the best of my knowledge, the data submitted here are not subject to any restrictions that would prevent their open sharing through the databases operated by the BioData Science Initiative (BSI). I have also reviewed and agreed to the BSI Terms of Use, including the provisions on the rights and duties of submitters, the handling of human-derived research data, and the access and benefit-sharing (ABS) of genetic resources.

Read more:

- BSI Terms of Use: <https://www.ddbj.nig.ac.jp/policies-e.html>
- Applicable national and international laws and rules on access and benefit-sharing: <https://www.ddbj.nig.ac.jp/faq/en/abs-data-sharing-e.html>

Submit

利用規約（生物多様性や ABS を含む）に対する同意チェックボックスを登録システムに設置（2026年7月）

94

DDBJ の対応③：東南アジア各国の支援

第11回 D-STEP 講習会 with Japan-Cambodia Workshop 2025/12/16 講習会 第11回 D-STEP 講習会 with Japan-Cambodia Workshop
第44回 DDBJing 講習会 with 富山大学 2025/09/29 講習会 第44回 DDBJing 講習会 with 富山大学
第10回 D-STEP 講習会 with Japan-the Philippines Workshop 2025/01/22 講習会 第10回 D-STEP 講習会 with Japan-the Philippines Workshop
第43回 DDBJing 講習会 with 藤田医科大学 2024/06/24 講習会 第43回 DDBJing 講習会 with 藤田医科大学
第9回 D-STEP 講習会 with Japan-Malaysia Workshop 2024/02/19 講習会 第9回 D-STEP 講習会 with Japan-Malaysia Workshop
第8回 D-STEP 講習会 with Japan-Thailand Workshop 2023/01/19 講習会 第8回 D-STEP 講習会 with Japan-Thailand Workshop

金銭的な利益配分は営利企業拠出の基金で賄う方向
INSDC はバイオインフォマティクスやデータベース分野で技術支援

INSDC は新たなステージへ

- 増大するデータ vs 効率化・自動化 (AI)
- メンバーシップの拡大
- 高まる重要性和社会的責任