

# 次世代シーケンスデータベースを 使って公開NGSデータの検索と自分 のNGSデータを登録する

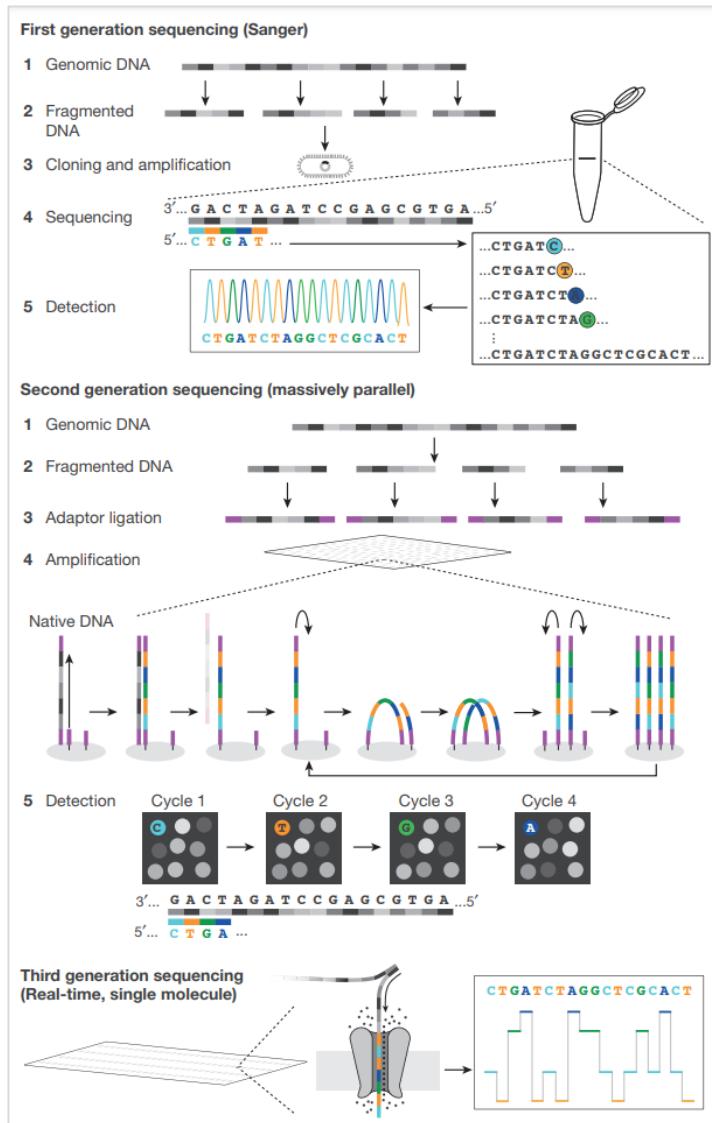
福田 亜沙美

国立遺伝学研究所 生命情報・DDBJセンター  
キュレータ

1. 次世代シーケンスデータベース
2. Sequence Read Archive
3. NGSデータの検索
4. NGSデータの DRA への登録

# 次世代シーケンスデータベース

# DNA シーケンサー



## 第一世代：サンガー法

- アプライドバイオシステム(ABI)

これ以降を次世代シーケンサーと呼ぶ

## 第二世代：超大量並列(short read)

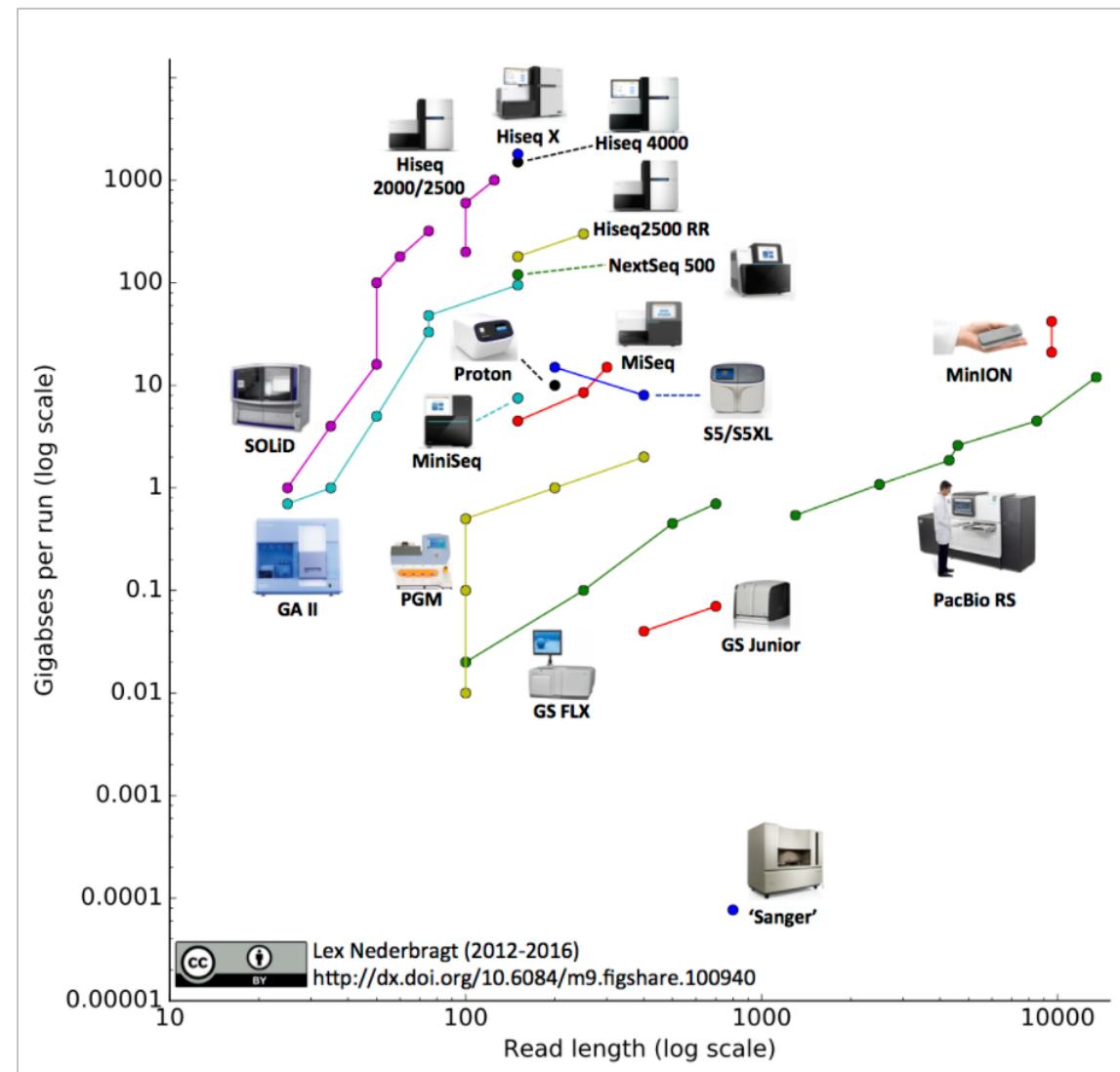
- Illumina

## 第三世代：リアルタイム一分子(long read)

- PacBio
- Oxford Nanopore

Shendure, J., Balasubramanian, S., Church, G. et al.  
Nature 550, 345–353 (2017). <https://doi.org/10.1038/nature24286>/  
Figure1

# 主な次世代シーケンサー



- Short read はさらに大量出力

Illumina NovaSeq 6000  
最大出力 6000 Gb

- Long read はさらに長く

Oxford Nanopore  
MinION < PromethION  
最長リードは1Mb以上

PacBio  
Sequel < Sequel II  
最長リードは175kb 以上

<https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/>

# 出力データと解析データ

生データ

リード

Quality value



fastq

```
@seq1
GGGTGATGGCCGCT...
+
IIIIII9IG9IC7...
```

アライメント

bam

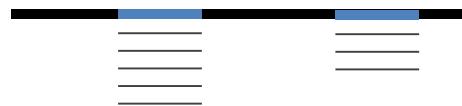
```
@HD VN:1.0 GO:none SO:coordinate
@SQ SN:chr1 LN:249698942
@SQ SN:chr2 LN:242508799
@SQ SN:chr3 LN:198450956
```

アセンブル/アノテーション付き配列データ

アセンブリ

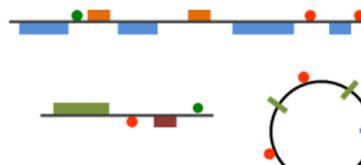


機能ゲノミクスデータ



- ・発現解析データ  
FPKM/RPKM, TPM etc.
- ・ChIP-Seq ピークファイル etc.

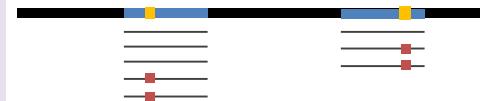
アノテーション



fasta

```
>seq1
GGGTGATGGCCGCT...
//
```

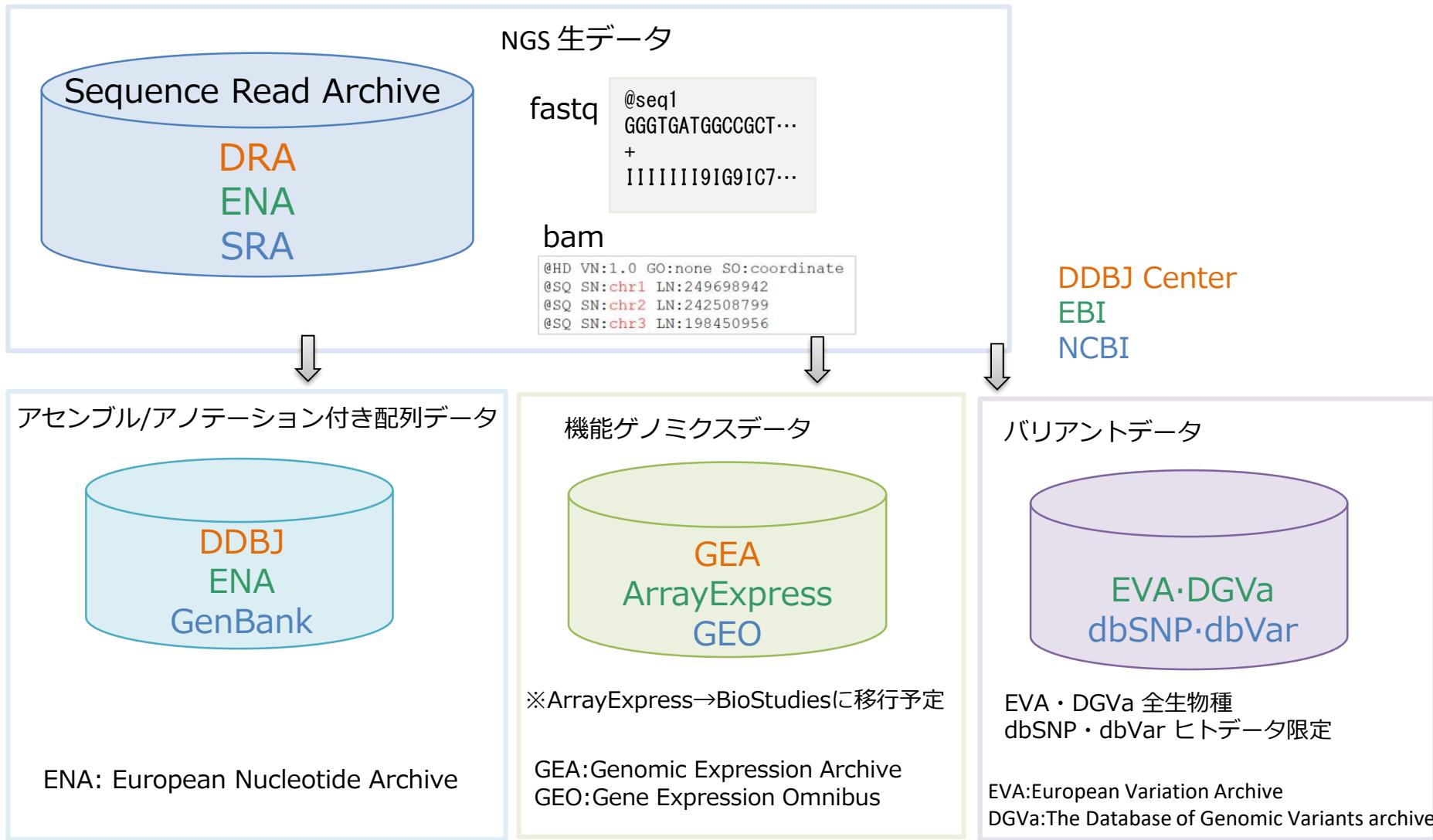
バリエントデータ



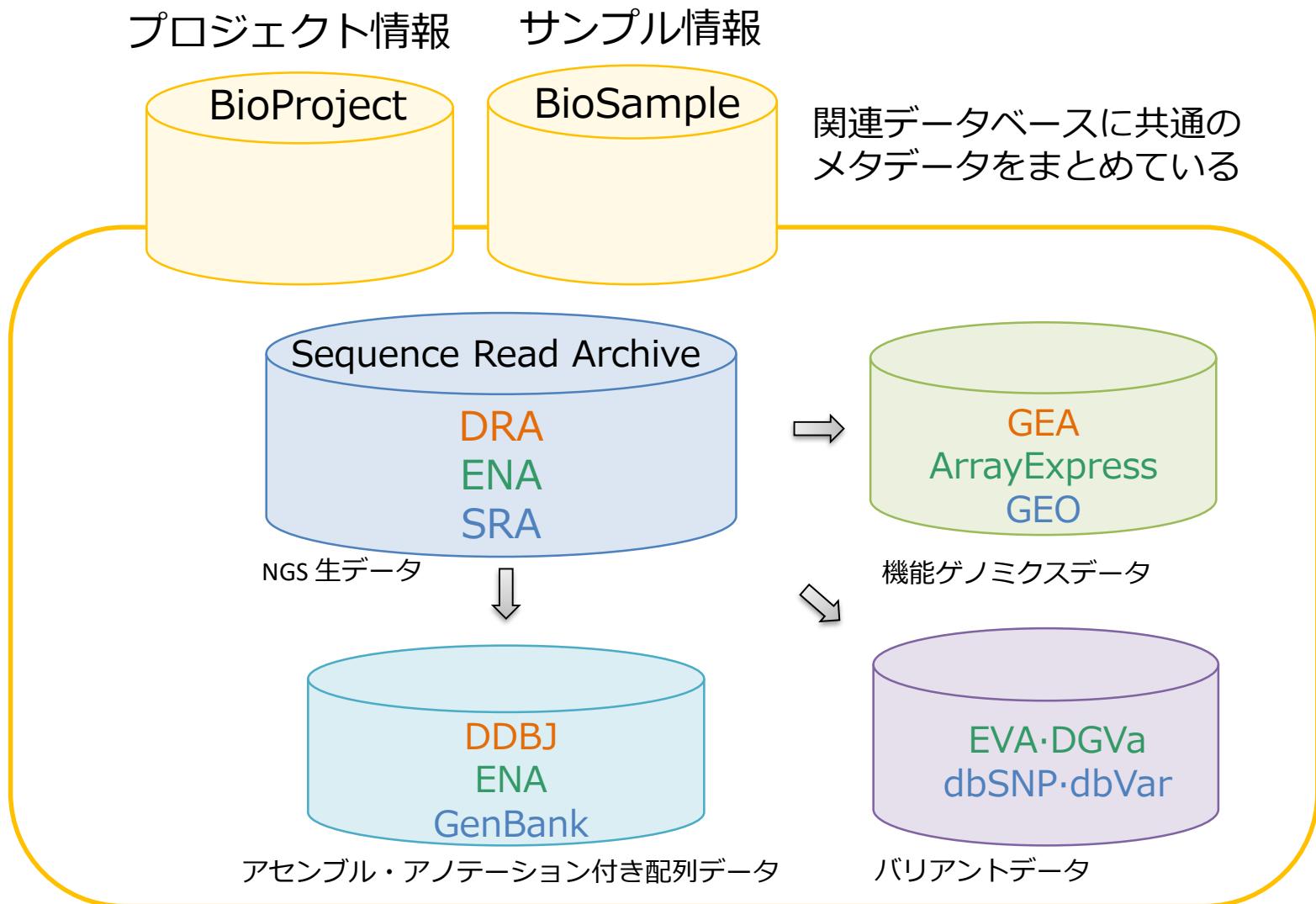
vcf

```
#CHROM POS ID REF ALT QUAL FILTER
20 14370 rs6054257 G A 29 PASS
20 17330 . T A 3 q10
20 1110696 rs6040355 A G, T 67 PASS
```

# NGSデータと対応するデータベース



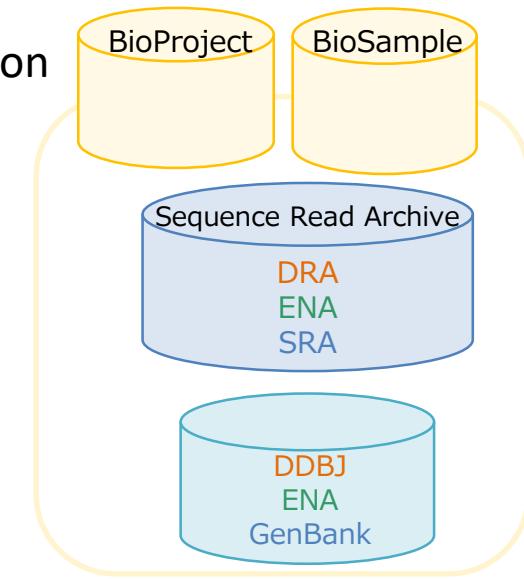
# データを整理するデータベース



※メタデータ: シーケンスデータを説明するデータ

## 国際塩基配列データベース

INSDC; International Nucleotide Sequence Database Collaboration



Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	<a href="#">Sequence Read Archive</a>		<a href="#">Sequence Read Archive</a>
Capillary reads	<a href="#">Trace Archive</a>		<a href="#">Trace Archive</a>
Annotated sequences	<a href="#">DDBJ</a>		<a href="#">GenBank</a>
Samples	<a href="#">BioSample</a>		<a href="#">BioSample</a>
Studies	<a href="#">BioProject</a>		<a href="#">BioProject</a>

<http://www.insdc.org/>

# アクセス制限データベース

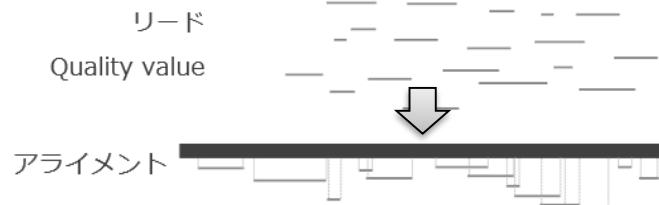
個人を特定可能な情報を含むヒトデータとそれ以外

↓  
アクセス制限

↓  
オープンアクセス

Japanese Genotype-phenotype Archive (JGA) ※NBDCと共に運営  
European Genome-Phenome Archive (EGA)  
The database of Genotypes and Phenotypes (dbGAP)

生データ



変異データ



機能ゲノミクスデータ



アセンブル/アノテーション付き配列データ



アノテーション



# 一次データベースと二次データベース

	一次データベース Primary database	二次データベース Secondary database
別の呼び方	Archival database	Curated database; Knowledgebase
データソース	研究者（登録者）が実験で得たデータを直接登録	一次データベースのデータや文献を解析、解釈、キュレーションした結果
例	<ul style="list-style-type: none"><li>•DRA/ERA/SRA (INSDC)</li><li>•DDBJ/ENA/GenBank (INSDC)</li><li>•GEA/ArrayExpress/GEO</li><li>•EVA·DGVa/dbSNP·dbVar</li><li>•PDB</li></ul>	<ul style="list-style-type: none"><li>•RefSeq</li><li>•Ensembl</li><li>•Expression Atlas</li><li>•ChIP-Atlas</li><li>•UniProt</li></ul>

# Sequence Read Archive

# データ量(NCBI SRA)

SRA data は 2019 年時点で 36PB, 2023年には 43 PB になる見込み

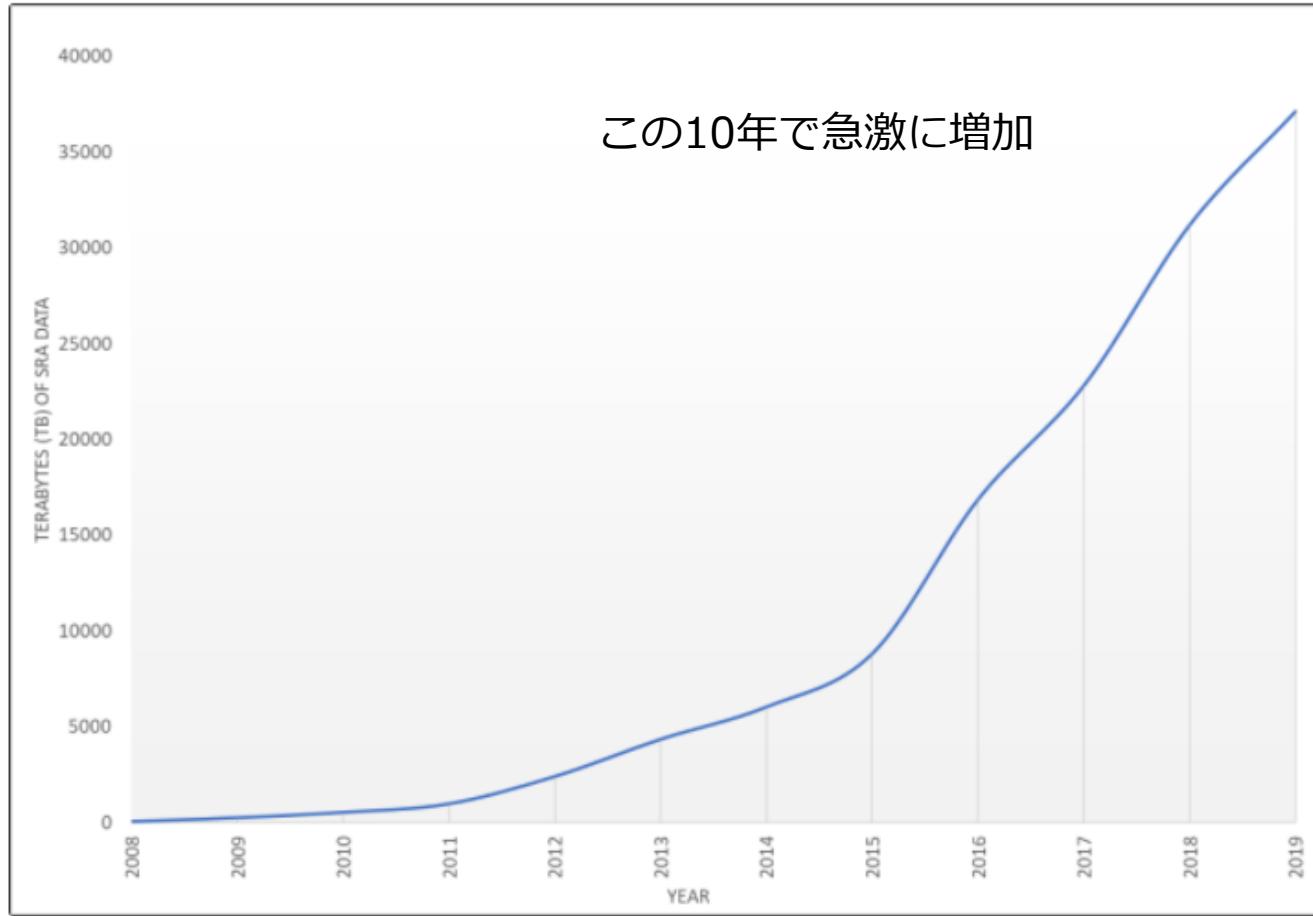


Figure 1. SRA data has grown exponentially over the last decade.

<https://ncbiinsights.ncbi.nlm.nih.gov/2020/06/30/sra-rfi/>

# 公開データフォーマット

fastq

```
@seq1
GGGTGATGGCCGCT...
+
IIIIIIII9IG9IC7...
```

bam

```
@HD VN:1.0 GO:none SO:coordinate
@SQ SN:chr1 LN:249698942
@SQ SN:chr2 LN:242508799
@SQ SN:chr3 LN:198450956
```

シーケンサー出力  
(PacBio hdf5 etc.)

SRA Toolkit →



アーカイブ用 SRA ファイル

SRA Toolkit →



fastq

ENA

NCBI

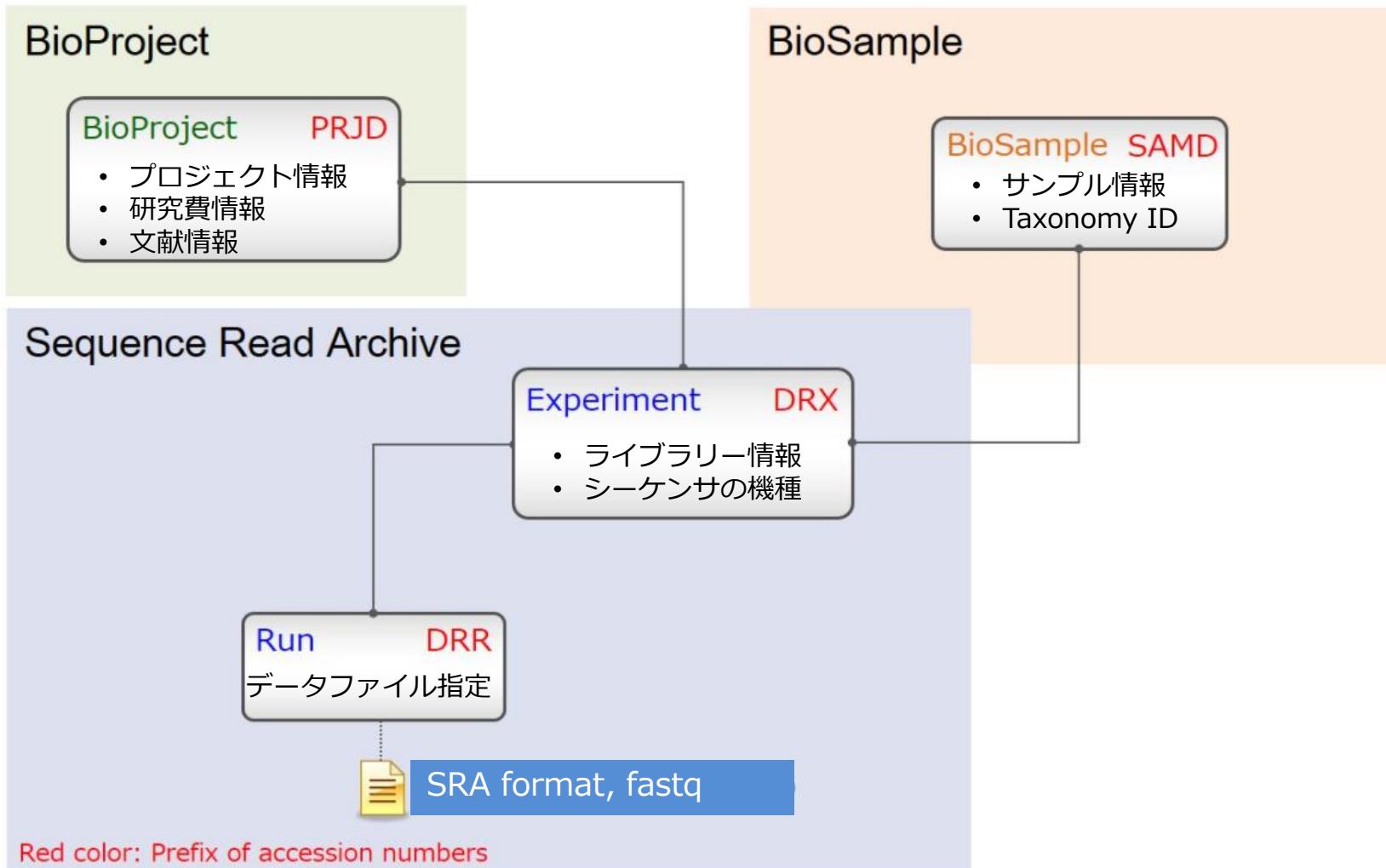
データ交換

公開データフォーマット

DDBJ	ENA	NCBI
sra, fastq	sra, fastq	sra

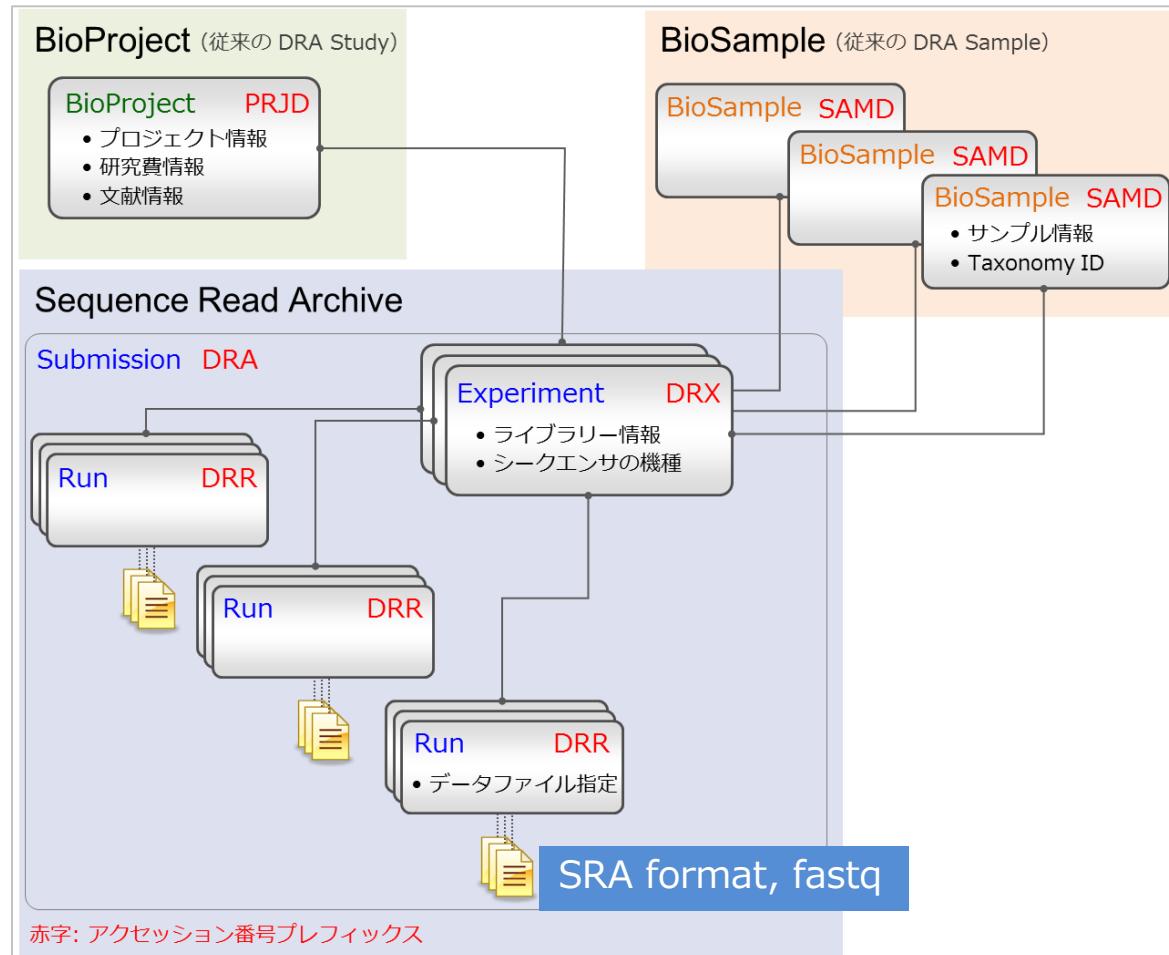
SRA format には配列データしか含まれておらず、  
メタデータは含まれていない

# SRAメタデータ



✓ SRA データは3つのデータベースにまたがる

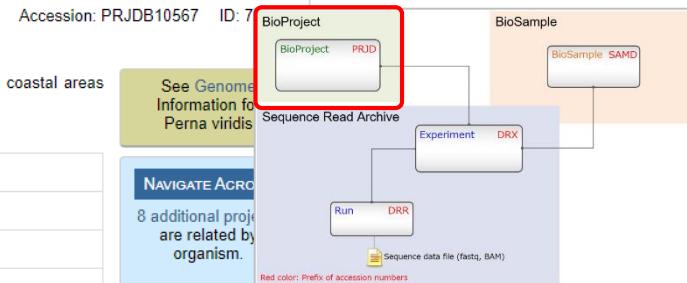
1つのプロジェクトに複数の Sample, Experiment, Run が関連付けられる



プロジェクト概要

Perna viridis (Asian green mussel)  
Asian green mussel whole genome sequencing  
The Asian green mussel, Perna viridis, is a bivalve species that is dominant in tropical and subtropical coastal areas around Asia. More...

Accession	PRJDB10567
Data Type	Genome sequencing and assembly
Scope	Monoisolate
Organism	Perna viridis [Taxonomy ID: 73031] Eukaryota; Metazoa; Spiralia; Lophotrochozoa; Mollusca; Bivalvia; Autobranchia; Pteriomorphia; Mytiloidea; Mytilidae; Mytilinae; Perna; Perna viridis
Publications	Inoue K <i>et al.</i> , "Genomics and Transcriptomics of the green mussel explain the durability of its byssus.", <i>Sci Rep</i> , 2021 Mar 16;11(1):5992
Grants	"Platform for Advanced Genome Science" (Grant ID 16H06279, Japan Society for the Promotion of Science)
Submission	Registration date: 25-Mar-2021 Atmosphere and Ocean research Institute, The University of Tokyo
Relevance	Evolution



アクセスション番号プレフィックス  
**PRJD (DDBJ)**  
**PRJE (ENA)**  
**PRJN (NCBI)**

関連データ

SRA データ

BioSample  
Assembly

SRA  
データサイズ

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	12
PUBLICATIONS	
PubMed	1
PMC	1
OTHER DATASETS	
BioSample	8
Assembly	1

Assembly details:

Assembly	Level	WGS	BioSample	Taxonomy
GCA_018327765.1	Scaffold	BNGV00000000	SAMD00247165	Perna viridis

SRA Data Details

Parameter	Value
Data volume, Gbases	125
Data volume, Mbytes	51677

- ✓ プロジェクトに関する情報
- ✓ 関連データをトップレベルで整理・ナビゲーションの入口

タイトル →

**Wild Perna viridis from Enoshima**

生物情報 →

Identifiers BioSample: SAMD00247165; SRA: DRS176506

パッケージ →

Organism [Perna viridis \(Asian green mussel\)](#)

cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Spiralia; Lophotrochozoa; Mollusca; Bivalvia; Pteriomorpha; Mytiloidea; Mytiloidea; Mytilidae; Mytilinae; Perna

Package [MIGS: eukaryote; version 5.0](#)

サンプル属性

Attributes

sample name	Perna_viridis_enoshima
collection date	2018-08-27
broad-scale environmental context	coastal sea
local-scale environmental context	intertidal zone
environmental medium	rock
estimated size	726Mb
geographic location	<a href="#">Japan Kanagawa Enoshima</a>
isolation and growth condition	<a href="https://www.jstage.jst.go.jp/article/sosj1996/21/1/21_1_19/_article/-char/ja/">https://www.jstage.jst.go.jp/article/sosj1996/21/1/21_1_19/_article/-char/ja/</a>
latitude and longitude	35.18 N 139.28 E
number of replicons	missing
ploidy	diploid
project name	Asian green mussel whole genome sequence
propagation	sexual
cultivar	missing
ecotype	missing
isolate	missing
strain	missing

Description Keywords: GSC:MiS;MIGS:5.0

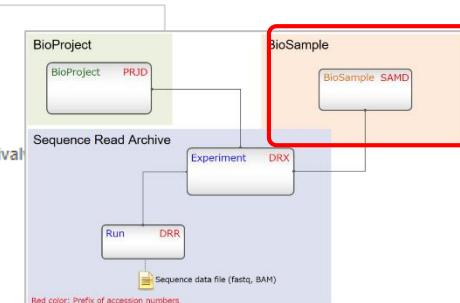
BioProject [PRJDB10567 Perna viridis](#)  
[Retrieve all samples from this project](#)

Submission [Atmosphere and Ocean research Institute, The University of Tokyo](#); 2021-03-25

Accession: SAMD00247165 ID: 18491400

[BioProject](#) [SRA](#) [Nucleotide](#)

関連データ →



アクセシション番号プレフィックス

SAMD (DDBJ)

SAME (ENA)

SAMN (NCBI)

- ✓ データを取得したサンプルが「属性名:値」のペアで記述されている
- ✓ サンプルの種類毎のパッケージ（必須・任意属性セット）

NCBI BioSample: <https://www.ncbi.nlm.nih.gov/biosample/SAMD00247165>

AJACS オンライン7

# SRA Experiment

[DRX234771](#): Illumina HiSeq 2500 paired end sequencing of SAMD00247165  
1 ILLUMINA (Illumina HiSeq 2500) run: 39.4M spots, 7.9G bases, 2.9Gb downl.

BioProject →

Submitted by: AORI

Study: Asian green mussel whole genome sequencing  
[PRJDB10567](#) • [DRP007082](#) • [All experiments](#) • [All runs](#)  
[show Abstract](#)

BioSample →

Sample: Wild Perna viridis from Enoshima

[SAMD00247165](#) • [DRS176506](#) • [All experiments](#) • [All runs](#)

Organism: [Perna viridis](#)

ライブラリー → Library:

Instrument: Illumina HiSeq 2500

Strategy: WGS

Source: GENOMIC

Selection: RANDOM

Layout: PAIRED

Construction protocol:

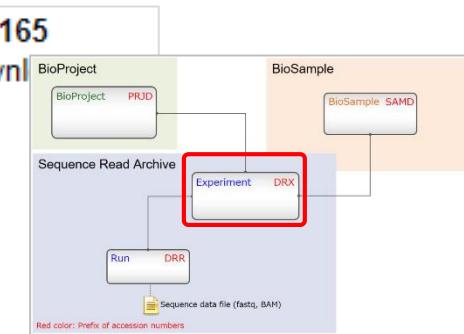
Spot descriptor:



Run →

Runs: 1 run, 39.4M spots, 7.9G bases, [2.9Gb](#)

Run	# of Spots	# of Bases	Size	Published
<a href="#">DRR244978</a>	39,412,069	7.9G	2.9Gb	2021-03-25



アクセスション番号プレフィックス  
DRX (DDBJ)  
ERX (ENA)  
SRX (NCBI)

- ✓ ライブラリー + シークエンサー
- ✓ BioProject, BioSample と Run をつなぐ  
中心オブジェクト

NCBI Experiment: <https://www.ncbi.nlm.nih.gov/sra/DRX234771>

# SRA Run

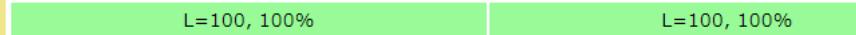
## Illumina HiSeq 2500 paired end sequencing of SAMD00247165 (DRR244978)

Metadata Analysis Reads Data access

Run	Spots	Bases	Size	GC content	Published	Access Type
DRR244978	39.4M	7.9Gbp	3.1G	34.6%	2021-03-25	public

Quality graph ([bigger](#))

This run has 2 reads per spot:



Legend

Metadata Analysis Reads Data access

Filter:  Find Filtered Download [What does it mean?](#)

[What can the filter be applied to?](#)

The Run is too big (>1.1G) for searching by sequence substring

Experiment DRX234771

Biosample

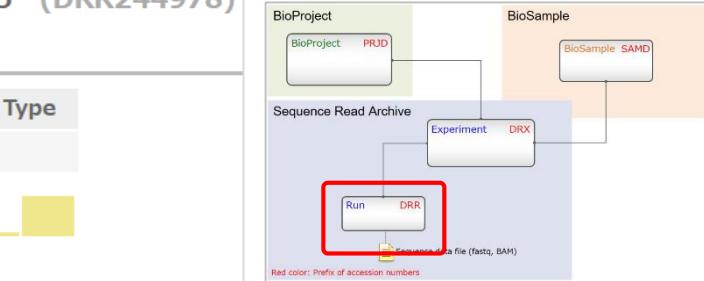
SAMD00247165

Bioproject

PRJDB10567

Show abstract

リード



アクセスション番号プレフィックス  
DRR (DDBJ)  
ERR (ENA)  
SRR (NCBI)

- ✓ リードデータのコンテナー
- ✓ リードには連番をアサイン

Reads (separated)

1. DRR244978.1 DRS176506

name: HWI-D00440:389:HFYCKBCX2:1:1101:1127:2068

member: default

x: 1127, y: 2068

>gnl|SRA|DRR244978.1.1 HWI-D00440:389:HFYCKBCX2:1:1101:1127:2068 (Biological)  
CAGGTGGTCGTGCGACACATTAATAAATATAATAGGAGATCTGTATAAGAGACAGTAATAA  
ATATGATTCTGTTTTTTGTTGATTTCATTGTGGTGTCTAA

One channel quality score

35 35 35 35 35 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40  
40  
40  
40 40

2. DRR244978.2 DRS176506

name: HWI-D00440:389:HFYCKBCX2:1:1101:1071:2073

member: default

x: 1071, y: 2073

>gnl|SRA|DRR244978.1.2 HWI-D00440:389:HFYCKBCX2:1:1101:1127:2068 (Biological)  
CAGGAAGCAACAGATGGAACATTACCATCTCTACAATACACAAGACCAAAAAACTAGTC  
AAATTATAACAACAAAACAACTAAAATGAAACACAACAAAC

One channel quality score

35 35 33 35 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40  
40  
40  
40 40

3. DRR244978.3 DRS176506

name: HWI-D00440:389:HFYCKBCX2:1:1101:1084:2120

member: default

x: 1084, y: 2120

35 35 33 35 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40  
40  
40  
40 40

NCBI Run: <https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=DRR244978>

# NGSデータを検索する

DRASearch

ENA Browser

NCBI SRA Entrez

# 検索方法

NGS データ検索は主にメタデータによる検索

- アクセッション番号で検索
- Keywordで検索
  - 論理演算子(AND, OR, NOT)の使用
  - ワイルドカード (\*)

Element	Meaning	Usage	Example
AND	In addition to	term1 AND term2	<i>glutathione AND transferase</i>
OR	Equivalence	term1 OR term2	<i>glutathione OR transferase</i>
NOT	Exclusion	term1 NOT term2	<i>coding NOT fragment</i>
*	Wildcard	partialTerm*	<i>gluta*</i>
" "	Exact match	"quoted text"	<i>"x-ray diffraction"</i>
( )	Grouping	(text)	<i>(reductase OR transferase) AND glutathione</i>
Field:	Field-specific search	fieldId:term	<i>description:dopamine</i>

[https://www.ebi.ac.uk/ebisearch/documentation.ebi#query\\_syntax](https://www.ebi.ac.uk/ebisearch/documentation.ebi#query_syntax)

対象データは DRA/ERA/SRA



The screenshot shows the DRASearch interface. In the search bar, the accession number "DRR000003" is entered. Below the search bar, there are filters for Organism, StudyType, Platform, and CenterName. An orange arrow points from the search bar to the search results area. The results are titled "DRR000003で検索した結果". The "Run Detail" section displays the following information:

Alias	DRR000003
Instrument model	
Date of run	2009-03-26
Run center	UT-MGS
Number of spots	6,459,732
Number of bases	232,550,352

The "READS (joined)" section shows the first few lines of sequence data:

```
>DRR000003.1
NGGCCAGGTGCACAAGCTCCTACCCCTGTGCTCACCA
>DRR000003.2
NAGATCCTGCCAGAAATTGTGTTCACTCTTCTTTT
>DRR000003.3
NGTACCAATCGCTGACTATGACAGGGAGGCTCACACG
>DRR000003.4
NTCTCACAAACTGGATTGTGATGTTGTCTCACT
```

The "Navigation" section provides links to related data:

- Submission: DRA000003 (with an FTP link)
- Study: DRP000003
- Experiment: DRX000003
- Sample: DRS000003 (with FASTQ and SRA links)

A callout box highlights the "FTP" link under the Submission entry.

fastq, SRA file の  
ftp サイトへのリンク

ダウンロードの方法は以下を参照

<https://www.ddbj.nig.ac.jp/faq/ja/how-to-download-data.html>

# DRASearch – keyword 検索

## メタデータ XML 上を検索

DRASearch

Accession :

Organism :  StudyType :

CenterName :  Platform :

Keyword : covid-19 OR sars-cov-2 OR (2019 AND ncov)

Show 20 records Sort by Study Search Clear

Keyword で検索  
※ワイルドカードは使用不可

Search Results ( 1136611 records ) Filter で絞り込み << < 1 / 56

Filtered by

document type:submission(556395) sample(203426) experiment(188770) run(187617) study(403)  
organism:Severe acute respiratory syndrome coronavirus 2(773364) Homo sapiens(318) Mus musculus(33) human gut metagenome(27)  
human nasopharyngeal metagenome(25) human metagenome(18)

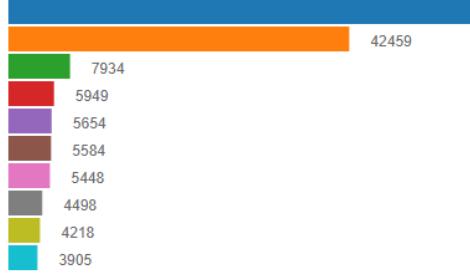
#	META_FILE	ACCESSION	STUDY	STUDY_TITLE	STUDY_TYPE	ORGANISM	BASES	SU
1	<a href="#">SRA1078496.study.xml</a> platelets take-up <b>SARS-COV2</b> mRNA independent of ACE2. Resting platelets from <b>COVID-19</b> patients had increased	<a href="#">SRP262885</a>	<a href="#">SRP262885</a>	RNA-seq of platelets from SARS-CoV-2 Covid-19	<u>Other</u>	<u>Homo sapiens</u>	273.8G	
2	<a href="#">SRA1066962.study.xml</a> </EXTERNAL_ID> </IDENTIFIERS> <DESCRIPTOR> <STUDY_TITLE> <b>SARS-Cov-2</b> virus strain ISR_JP0320	<a href="#">SRP257463</a>	<a href="#">SRP257463</a>	SARS-Cov-2 virus strain ISR_JP0320 genome sequencing	<u>Other</u>	<u>Severe acute respiratory syndrome coronavirus 2</u>	446.6M	
3	<a href="#">SRA1197347.study.xml</a> during <b>COVID19</b> . Overall design: Comparison of hospitalized subjects: 30 <b>Sars-COV2</b> -negative versus 65 <b>Sars-COV</b>	<a href="#">SRP306910</a>	<a href="#">SRP306910</a>	PolyA RNA-seq from whole blood of Sars-COV2-negative and -positive subjects	<u>Transcriptome Analysis</u>	<u>Homo sapiens</u>	0	

DBCLS Research Services Contact About  
 DRA Home  DDBJ flat file search

 DBCLS SRA > SRA ✓ SRA BioProject BioSample

Keyword :   
 Accession :  アクセッショント番号で検索

Show 25 records Sort by Updated Order DESC

BioProject OrganismName		Details for DRR000001	DRR000001を検索した結果	JSON																																								
 Homo sapiens	 Sorghum bicolor																																											
 Mus musculus	 Panicum virgatum																																											
 soil metagenome	 Oryza sativa																																											
 Populus trichocarpa	 Zea mays																																											
 Arabidopsis thaliana	 Rattus norvegicus																																											
 42459 7934 5949 5654 5584 5448 4498 4218 3905		<b>DRR000001を検索した結果</b> <b>BioProject: PRJDA38027</b> <table border="1"> <tr><td>Title</td><td>Bacillus subtilis subsp. natto BEST195 genome sequencing project</td></tr> <tr><td>Description</td><td>&lt;p&gt; &lt;b&gt;&lt;i&gt;Bacillus subtilis&lt;/i&gt;&lt;/b&gt; subsp. &lt;i&gt;natto&lt;/i&gt; BEST195&lt;/b&gt;. &lt;i&gt;Bacillus subtilis&lt;/i&gt; subsp. &lt;i&gt;natto&lt;/i&gt; BEST195 was isolated from fermented soybeans and will be used for comparative genome analysis.</td></tr> <tr><td>Organism name</td><td>Bacillus subtilis subsp. natto BEST195</td></tr> <tr><td>Archive</td><td>DDBJ</td></tr> <tr><td>LocusTagPrefix</td><td>BSNT</td></tr> <tr><td>Organization name</td><td>Keio University</td></tr> <tr><td>Submitted</td><td>2009-05-13</td></tr> <tr><td>Publication ID</td><td>25329997</td></tr> <tr><td>Data Type</td><td>Genome sequencing</td></tr> </table> <b>STUDY</b> <table border="1"> <tr><td>Study Title</td><td>Bacillus subtilis subsp. natto BEST195 genome sequencing project</td></tr> <tr><td>Abstract</td><td>&lt;b&gt;&lt;i&gt;Bacillus subtilis&lt;/i&gt;&lt;/b&gt; subsp. &lt;i&gt;natto&lt;/i&gt; BEST195&lt;/b&gt;. &lt;i&gt;Bacillus subtilis&lt;/i&gt; subsp. &lt;i&gt;natto&lt;/i&gt; BEST195 was isolated from fermented soybeans and will be used for comparative genome analysis.</td></tr> <tr><td>Study Type</td><td>Whole Genome Sequencing</td></tr> <tr><td>Center Name</td><td>KEIO</td></tr> <tr><td>Center Project Name</td><td>Bacillus subtilis subsp. natto BEST195</td></tr> <tr><td>XREF ID</td><td>25329997</td></tr> </table> <b>EXPERIMENT, RUN and BioSample</b> <table border="1"> <tr><td>EXPERIMENT: DRX000001</td><td></td></tr> <tr><td>Title</td><td>B. subtilis subsp. natto genome sequencing September 2008</td></tr> <tr><td>Instrument Model</td><td>Illumina Genome Analyzer II</td></tr> <tr><td>RUN</td><td></td></tr> <tr><td>DRR000001</td><td>  sra fastq</td></tr> </table>			Title	Bacillus subtilis subsp. natto BEST195 genome sequencing project	Description	<p> <b><i>Bacillus subtilis</i></b> subsp. <i>natto</i> BEST195</b>. <i>Bacillus subtilis</i> subsp. <i>natto</i> BEST195 was isolated from fermented soybeans and will be used for comparative genome analysis.	Organism name	Bacillus subtilis subsp. natto BEST195	Archive	DDBJ	LocusTagPrefix	BSNT	Organization name	Keio University	Submitted	2009-05-13	Publication ID	25329997	Data Type	Genome sequencing	Study Title	Bacillus subtilis subsp. natto BEST195 genome sequencing project	Abstract	<b><i>Bacillus subtilis</i></b> subsp. <i>natto</i> BEST195</b>. <i>Bacillus subtilis</i> subsp. <i>natto</i> BEST195 was isolated from fermented soybeans and will be used for comparative genome analysis.	Study Type	Whole Genome Sequencing	Center Name	KEIO	Center Project Name	Bacillus subtilis subsp. natto BEST195	XREF ID	25329997	EXPERIMENT: DRX000001		Title	B. subtilis subsp. natto genome sequencing September 2008	Instrument Model	Illumina Genome Analyzer II	RUN		DRR000001	  sra fastq
Title	Bacillus subtilis subsp. natto BEST195 genome sequencing project																																											
Description	<p> <b><i>Bacillus subtilis</i></b> subsp. <i>natto</i> BEST195</b>. <i>Bacillus subtilis</i> subsp. <i>natto</i> BEST195 was isolated from fermented soybeans and will be used for comparative genome analysis.																																											
Organism name	Bacillus subtilis subsp. natto BEST195																																											
Archive	DDBJ																																											
LocusTagPrefix	BSNT																																											
Organization name	Keio University																																											
Submitted	2009-05-13																																											
Publication ID	25329997																																											
Data Type	Genome sequencing																																											
Study Title	Bacillus subtilis subsp. natto BEST195 genome sequencing project																																											
Abstract	<b><i>Bacillus subtilis</i></b> subsp. <i>natto</i> BEST195</b>. <i>Bacillus subtilis</i> subsp. <i>natto</i> BEST195 was isolated from fermented soybeans and will be used for comparative genome analysis.																																											
Study Type	Whole Genome Sequencing																																											
Center Name	KEIO																																											
Center Project Name	Bacillus subtilis subsp. natto BEST195																																											
XREF ID	25329997																																											
EXPERIMENT: DRX000001																																												
Title	B. subtilis subsp. natto genome sequencing September 2008																																											
Instrument Model	Illumina Genome Analyzer II																																											
RUN																																												
DRR000001	  sra fastq																																											

Database Center for Life Science  
[Contact](#)  
[Site policy](#)  
 © 2021 DBCLS

fastq, SRA file の  
 DDBJ ftp サイトへのリンク

ENAS European Nucleotide Archive

Home Submit ▾ Search ▾ Rulespace About ▾ Support ▾

Enter text search terms  Search

Examples: histone, BN0000065

Enter accession  View

Examples: Taxon:9606, BN0000065, PRJEB402

## Searching ENA

ENAS data can be searched and retrieved interactively and programmatically and visualized using the ENA Browser. Please refer to the following sections for more information about the ENA data access functionality with links to more detailed documentation.

Search term:  Search

Uses EBI Search to perform a free text search across ENA data.

Free Text Search Advanced Search Cross References Sequence Similarity Search Sequence Version Archive

アクセスション番号で検索

対象データは Read のみではなく、  
ENA 全データを含む

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	<a href="#">Sequence Read Archive</a>	European Nucleotide Archive ( <a href="#">ENA</a> )	<a href="#">Sequence Read Archive</a>
Capillary reads	<a href="#">Trace Archive</a>		<a href="#">Trace Archive</a>
Annotated sequences	<a href="#">DDBJ</a>		<a href="#">GenBank</a>
Samples	<a href="#">BioSample</a>		<a href="#">BioSample</a>
Studies	<a href="#">BioProject</a>		<a href="#">BioProject</a>

2020年8月から新ブラウザへ完全移行

<https://www.ebi.ac.uk/about/news/service-news/retirement-old-ena-browser-5th-august-2020>

# ENA browser - 一覧表示

BioProjectでの  
検索結果

プロジェクト情報

オブジェクトの  
一覧表示

Project: PRJDB5174

PRJDB5174を検索した結果

Organism: flower metagenome

Secondary Study Accession: DRP003904

Study Title: Ohanami Metagenome

Center Name: Ohanami Metagenome Project

Study Name: flower metagenome

ENA-REFSEQ: N

PROJECT-ID: 419333

ENA-FIRST-PUBLIC: 2017-11-29

ENA-LAST-UPDATE: 2018-01-28

Show More

View: XML  
XML (STUDY)

Download: XML  
XML (STUDY)

Navigation: Show

Read Files: Hide

Related ENA Records: Show

Read Files

Show Column Selection

Download report: JSON TSV

Download Files as ZIP

Download selected files

Study Accession	Sample Accession	Experiment Accession	Run Accession	Tax Id	Scientific Name	Download All	FTP	Size
PRJDB5174	SAMD00062428	DRX067983	DRR074069	1385665	flower metagenome	<input type="checkbox"/> DRR074069_1.fastq.gz	<input type="checkbox"/> DRR074069_2.fastq.gz	
PRJDB5174	SAMD00062429	DRX067984	DRR074070	1385665	flower metagenome	<input type="checkbox"/> DRR074070_1.fastq.gz	<input type="checkbox"/> DRR074070_2.fastq.gz	

<https://www.ebi.ac.uk/ena/browser/view/PRJDB5174>

# ENA browser – keyword 検索

## Text Search

Uses EBI Search to perform a free text search across ENA data. For more detailed usage please refer to the help & documentation section.

### Search term:

covid-19 OR sars-cov-2 OR (2019 AND ncov)

Search

covid-19 OR sars-cov-2 OR (2019 AND ncov)

Search results for covid-19 OR sars-cov-2 OR (2019 AND ncov)

- Assembly
  - Assembly (705)
- Sequence
  - Sequence (549,340)
  - Sequence (Standard) (549,340)
- Contig set
  - Genome assembly contig set (19)
- Coding
  - Coding (3,325,502)
  - Coding (Standard) (3,312,684)
  - Coding (TLS) (12,818)
- Non-coding
  - Non-coding (2,691)

## Datatype で絞り込み

- Read
  - Experiment (457,788)
  - Run (84,582)
- Analysis
  - Analysis (420,698)
- Study
  - Study (747)
  - Project (916)

対象データは Read のみではなく、  
ENA 全データを含む

<b>Assembly</b> View all 705 results.	
GCA_015251995.1	Vero_WHO_a1.0 assembly for Chlorocebus sabaeus
<b>Sequence</b> View all 549,340 results.	
MW367318	Homo sapiens clone TAU-1109_LC_VJ anti-SARS-CoV-2 neutralizing immunoglobulin kappa light chain variable region mRNA, partial cds.
<b>Sequence (Standard)</b> View all 549,340 results.	
MW367318	Homo sapiens clone TAU-1109_LC_VJ anti-SARS-CoV-2 neutralizing immunoglobulin kappa light chain variable region mRNA, partial cds.
<b>Genome assembly contig set</b> View all 19 results.	
JACXSZ010000000	project.
<b>Coding</b> View all 3,325,502 results.	
QOS44862	synthetic construct SARS-CoV-2 nucleocapsid
<b>Coding (Standard)</b> View all 3,312,684 results.	
QOS44862	synthetic construct SARS-CoV-2 nucleocapsid
<b>Coding (TLS)</b> View all 12,818 results.	
NSM03644	Mus musculus (house mouse) partial immunoglobulin heavy chain junction region
<b>Non-coding</b> View all 2,691 results.	
MW187860.1:2169..2393:misc_RNA	synthetic construct antisense RNA Sok

[https://www.ebi.ac.uk/ena/browser/text-search?query=covid-19%20OR%20sars-cov-2%20OR%20\(2019%20AND%20ncov\)](https://www.ebi.ac.uk/ena/browser/text-search?query=covid-19%20OR%20sars-cov-2%20OR%20(2019%20AND%20ncov))

# ENA browser - Advanced Search

## ①-⑤の手順で段階的に絞り込む

**Advanced Search**

Customise your own search query and retrieve a set of ENA records tailored to your search criteria.

All searches are performed against a subset of the archive specified by the *Data type* you choose to search against. You can then build your search query to specify what data you are looking for and select what fields you want to retrieve from your search. There are additional options to include/exclude specific datasets as well as filter the number of results you wish to return.

If you intend to repeat the same search at a later date, you can save this as a Rule using Rulespace. If you want to access the same data programmatically, you can copy the produced curl request and run this yourself against the ENA Portal API.

1 Data Type    2 Query    3 Inclusion/Exclusion    4 Fields    5 Data Filters

Select datatype to build a new search

Data type: Raw reads

NGS データ検索 = Raw reads

Or

Use a previously defined Rulespace search

Rule ID/Name:

Load Rule    Reset

Next ▶

<https://www.ebi.ac.uk/ena/browser/advanced-search>

使い方のドキュメント

<https://ena-docs.readthedocs.io/en/latest/retrieval/advanced-search.html>

# Advanced Search - Query

## クエリを細かく指定

項目がカテゴリごとに表示され、様々な項目で高度な絞り込みができる  
例1) collection\_date が 2011-01-01 以降  
例2) 地図上から位置を絞り込む

クエリの指定

The screenshot shows the DDBJ Advanced Search interface. The top navigation bar has tabs: Data Type, Query (selected), Inclusion/Exclusion, Fields, Data Filters, and Results. The 'Query' tab is highlighted with a blue box. The main content area is titled 'クエリの指定' (Specify Query). On the left, a sidebar lists project categories: Taxonomy and related, Geographical location, Geography, Collection event information, Sampling information, Sample state and conditions, Host information, Methodology, Sequencing information, Database record, and File information. The 'Geographical location' category is selected. The main query builder shows a condition 'Collection date' set to '2020-01-08'. Below this is a map search interface for 'Geographical location' using a bounded box. The map shows Europe with a red box drawn around the United Kingdom. The map interface includes fields for 'Southwest point' (Latitude: 49, Longitude: -8) and 'Northeast point' (Latitude: 60, Longitude: 4).

# Advanced Search – Fields

## 出力項目を選択

1 Data Type      2 Query      3 Inclusion/Exclusion      4 Fields      5 Data Filters      6 Results

By default, your search will return the run accession and description/title for all results that match your query.

Default fields       Manually select fields

**Fields:** ecotype,fastq\_bytes,fastq\_ftp,fastq\_md5,collection\_date,fastq\_aspera,fastq\_galaxy      **Reset**

**Field sets:** FASTQ Files      SRA Files      Submitted Files      ?

**Select and order Fields:** ?

**Available Fields**

- cell\_line
- cell\_type
- center\_name
- checklist
- collected\_by
- collection\_date\_submitted
- completeness\_score
- contamination\_score
- country
- cram\_index\_aspera
- cram\_index\_ftp

**Selected Fields**

- ecotype
- fastq\_bytes
- fastq\_ftp
- fastq\_md5
- collection\_date
- fastq\_aspera
- fastq\_galaxy

取得ファイルフォーマットを選択

Table の結果に含める field を選択

◀ Back      Copy Curl Request       Next ▶      **Search **

# Advanced Search – Results

結果を表形式で表示

The screenshot shows the 'Advanced Search – Results' interface. At the top, there are five filter tabs: 'Data Type', 'Query', 'Inclusion/Exclusion', 'Fields', and 'Data Filters'. The 'Results' tab on the far right has a blue border and a green circle with the number '6' indicating the count of results. Below the tabs are three download buttons: 'Download ENA records: XML' (with a question mark icon), 'Download selected files: Bundled ZIP Individually' (with a question mark icon), and 'Download report: JSON TSV' (with a question mark icon). A central message says '結果を XML, JSON, TSV 形式でダウンロード可能' (Results can be downloaded in XML, JSON, TSV format). The main area displays a table of search results:

Run Accession	Sample Accession	Ecotype	FASTQ Bytes	Download All		Collection Date	FASTQ Aspera
				FASTQ FTP	FASTQ Md5		
DRR022337	SAMD00019161	Col-0	3GB 3GB	<input type="checkbox"/> DRR022337_1.fastq.gz <input type="checkbox"/> DRR022337_2.fastq.gz	77864c47a83cf6b0d2a1982e023dcab 79ab3fc34a12aad7b3af32519ec1dc6f	2011-01-01	fasp.sra.ebi.ac.uk:/vol fasp.sra.ebi.ac.uk:/vol
DRR022338	SAMD00019159	Col-0	3GB 3GB	<input type="checkbox"/> DRR022338_1.fastq.gz <input type="checkbox"/> DRR022338_2.fastq.gz	35b8c8f8c0683d3002ced64ec510663f dad839ac15a310336cc5ea478980d506	2011-01-01	fasp.sra.ebi.ac.uk:/vol fasp.sra.ebi.ac.uk:/vol
DRR022339	SAMD00019160	Col-0	3GB 3GB	<input type="checkbox"/> DRR022339_1.fastq.gz <input type="checkbox"/> DRR022339_2.fastq.gz	41103cb1884707b9e32422b32bc9cc6a 92e4efb0fdbdb8fa9c1235785be9f5dfd	2011-01-01	fasp.sra.ebi.ac.uk:/vol fasp.sra.ebi.ac.uk:/vol
DRR029379	SAMD00025039	Col-0	357MB	<input type="checkbox"/> DRR029379.fastq.gz	06950452b0f4f0321706fc49e0cbe794	2013-12-02	fasp.sra.ebi.ac.uk:/vol

対象データベースを SRA に指定

Query は Keyword, accession どちらも可

Experiment が公開情報の最小単位

# SRA 検索結果の詳細

参照Experimentの  
内容が表示される

BioProject →

BioSample →

ライブラリー  
シークエンサー →

Run →

The screenshot shows the SRA search results for experiment DRR148118. At the top, there is a search bar with "SRA" selected and "DRR148118" entered. Below the search bar are "Create alert" and "Advanced" buttons, and a "Search" button. To the right of the search bar are "Send to:" and "Help" buttons.

The main content area displays the following information:

- Full** (dropdown menu)
- DRX138869: Illumina HiSeq 2500 paired end sequencing of SAMD00135006**  
1 ILLUMINA (Illumina HiSeq 2500) run: 60.4M spots, 12.2G bases, 6.7Gb downloads
- Submitted by:** DBCLS
- Study:** RNA sequencing of Japanese stick insect
  - [PRJDB7316](#) • [DRP007253](#) • [All experiments](#) • [All runs](#)
  - [show Abstract](#)
- Sample:** midgut replicate1
  - [SAMD00135006](#) • [DRS183894](#) • [All experiments](#) • [All runs](#)
  - Organism:** [Entoria okinawaensis](#)
- Library:**
  - Instrument:** Illumina HiSeq 2500
  - Strategy:** RNA-Seq
  - Source:** TRANSCRIPTOMIC
  - Selection:** cDNA
  - Layout:** PAIRED
  - Construction protocol:**
- Spot descriptor:**
  - forward
  - reverse
- Runs:** 1 run, 60.4M spots, 12.2G bases, [6.7Gb](#)

Run	# of Spots	# of Bases	Size	Published
<a href="#">DRR148118</a>	60,433,221	12.2G	6.7Gb	2021-05-06

A callout box labeled "Run の詳細表示" points to the "Run" column of the table above. A sidebar on the right contains "Related information" links (BioProject, BioSample, PMC, PubMed, Taxonomy) and a section titled "関連情報へのリンク" with links to "論文情報" and "生物情報". The "Recent activity" section shows search history for DRR148118 (1) and DRR244978 (1). The URL <https://www.ncbi.nlm.nih.gov/sra/?term=DRR148118> is displayed at the bottom.

# NCBI SRA Entrez

Keyword で検索

検索結果は Experiment 単位で表示

covid-19 OR sars-cov-2 OR (2019 AND ncov)  
で検索した結果

様々な項目で絞り込み

アクセス制限  
or オープン

シーケンサー →

ライブラリーの構築手法

Access Controlled (732)  
Public (688,462)

Source DNA (34,475)  
RNA (705,146)

Type genome (19,693)

Library Layout paired (511,611)  
single (229,957)

Platform BGISEQ (1,684)  
Capillary (10)  
Illumina (638,074)  
Ion Torrent (2,926)  
Oxford Nanopore (98,801)  
PacBio SMRT (73)

Strategy EpiGenomics (46)  
Exome (16,881)  
Genome (35,848)  
RNASEq (54)  
other (688,739)

Data in Cloud GS (688,360)  
S3 (688,240)

File Type bam (191,253)  
cram (311,307)  
fastq (136,694)

Other aligned data (437,854)

[Clear all](#)

Summary 20 per page

Send to:

対象生物で絞り込み

Results by taxon

Top Organisms [Tree]  
Severe acute respiratory syndrome-related coronavirus (720451)  
Homo sapiens (10860)  
human gut metagenome (1399)  
metagenome (1081)  
human nasopharyngeal metagenome (1013)  
All other taxa (6764)

関連DBで絞り込み

Search in related databases

Database	Access		
	public	controlled	all
BioSample	719,323	770	720,093
BioProject	732	2	734
dbGaP		60	60
GEO Datasets	4,149		4,149

Find related data

Database: Select

Find items

Search details

(“Severe acute respiratory syndrome coronavirus 2”[Organism] OR covid-19[All Fields]) OR (“Severe acute respiratory syndrome coronavirus 2”[Organism] OR sars-cov-2[All Fields]) OR (2019[All Fields])

[https://www.ncbi.nlm.nih.gov/sra/?term=covid-19+OR+sars-cov-2+OR+\(2019+AND+ncov\)](https://www.ncbi.nlm.nih.gov/sra/?term=covid-19+OR+sars-cov-2+OR+(2019+AND+ncov))

# NCBI Advanced - SRA

Filter  
PubMed, PMC, Nucleotide, Assemblyなどと相互参照されているSRAレコードを検索できる  
例) "sra pmc"

Properties  
規定項目によって検索結果を絞り込む  
例) "instrument illumina novaseq 6000"

Builder

Organism: Homo sapiens  
Publication Date: 2019 to present  
Source: (genomic (5979758), genomic single cell (15944), metagenomic (2830632), metatranscriptomic (45174), other (146647), synthetic (81543), transcriptomic (2700407), transcriptomic single cell (90292), viral rna (770096))

Show index list (circled)  
Hide index list (circled)

Search or Add to history

History

Search	Add to builder	Query	Items found	Time
#2	Add	Search ("homo sapiens"[Organism]) AND "study type transcriptome analysis"[Properties]	617379	12:17:38
#1	Add	Search DRR000001	1	11:44:45

# NCBI Advanced – BioSample

The screenshot shows the NCBI BioSample Advanced Search Builder. At the top, there's a navigation bar with links for NCBI, Resources, and How To. Below it, a dropdown menu shows 'BioSample' and 'Advanced' (which is highlighted with an orange circle). A large orange arrow points from the 'Advanced' link down to the search interface.

The main search area has a heading 'BioSample Advanced Search Builder' with a thumbnail image of a tissue sample. The search query is displayed as '("strain"[Attribute Name]) AND BALB/c\*'. There's a 'Edit' link next to it.

A large blue callout box on the right contains the text 'BioSample の属性名、値で検索 例) strain BALB/c を含む'.

The search interface includes a 'Builder' section with a dropdown for 'Attribute Name' set to 'strain'[Attribute Name]. A scrollable list of attribute names and their counts is shown:

- strain (3027736)
- strain/age (22)
- strain/backgorund (22)
- strain/background (27743)
- strain/breed (256)
- strain/cell line (149)
- strain/cell line background (276)
- strain/cell type (252)
- strain/clone background (8)
- strain/cultivar (34)

Navigation links for the list are 'Previous 200' and 'Next 200'. A 'Refresh index' button is also present.

Below the builder, there are search parameters: 'AND' dropdown, 'All Fields' dropdown, and a search term 'BALB/c\*'. There are also buttons for 'Search' and 'Add to history'.

# NCBI Advanced – BioSample

SRAで関連データを探す

Find related data  
Database: SRA

Links to SRA experiments  
Find items

Search details

Experiment 単位で結果が表示される

Run Selector

View results as an expanded interactive table using the RunSelector. [Send results to Run selector](#)

Links from BioSample

Items: 1 to 20 of 1016

<< First < Prev Page 1 of 51 Next >> Last >>

Can only process the first 1000 items with links

GSM4836279\_ChIP\_CEBPb\_BALB\_BMDM\_24hL4\_rep2; Mus musculus; ChIP-Seq

1. 1 ILLUMINA (Illumina HiSeq 4000) run: 19.1M spots, 1.4G bases, 498.2Mb downloads  
Accession: SRX9311852

GSM4878919\_Main2\_010819.I2; Mus musculus; OTHER

2. 1 ILLUMINA (Illumina HiSeq 2500) run: 73.3M spots, 9.1G bases, 4Gb downloads  
Accession: SRX9427448

GSM4878918\_Main2\_010819.I1; Mus musculus; OTHER

3. 1 ILLUMINA (Illumina HiSeq 2500) run: 74.3M spots, 9.2G bases, 4.1Gb downloads  
Accession: SRX9427447

Results by taxon

Top Organisms [Tree]

- Mus musculus (965)
- mouse metagenome (26)
- mouse gut metagenome (10)
- Candida albicans (8)
- Pseudomonas lundensis (6)
- All other taxa (1)

More...

Find related data  
Database: Select

Find items

Recent activity

Turn Off Clear

# NCBI Run Selector

関連データ番号とBioSample, Experimentメタデータをまとめて表形式で表示

The screenshot shows the NCBI SRA Run Selector interface. At the top, there are navigation links for NCBI and SRA Run Selector, along with search, help, settings, and user icons.

**Filters List:**

- 1 AvgSpotLen
- 2 Bases
- 3 Bytes
- 4 rainfall\_regm

**Common Fields:**

BioProject	PRJDB5174
Consent	PUBLIC
Assay Type	AMPLICON
BioSampleModel	MIMARKS.survey.plant-associated
Center Name	DRCLS
DATASTORE filetype	SRA
DATASTORE provider	GS, NCBI, S3
DATASTORE region	gs.US, ncbi.public, s3.us-east-1

**共通情報** (Common Information) is annotated with a bracket pointing to the Common Fields section.

**メタデータをまとめてダウンロード** (Download Metadata) is annotated with a bracket pointing to the "Download" button in the "Select" section.

**表形式** (Table Format) is annotated with a bracket pointing to the results table.

**Results Table:**

	Run	BioSample	AvgSpotLen	Bases	Bytes	collection_date	Experiment	geo_loc_name	lat_lon
1	DRR074069	SAMD00062428	139	3.74 M	2.13 Mb	2015-03-31	DRX067983	Japan:Fukuoka, Kurume	33.1946 N 130.3211 E
2	DRR074070	SAMD00062429	151	4.36 M	2.61 Mb	2015-04-02	DRX067984	Japan:Fukuoka, Kurume	33.1946 N 130.3211 E
3	DRR074071	SAMD00062430	144	2.53 M	1.45 Mb	2015-03-31	DRX067985	Japan:Oita, Yufu	33.2100 N 131.5400 E
4	DRR074072	SAMD00062431	247	6.02 M	3.65 Mb	2015-03-31	DRX067986	Japan:Oita, Yufu	33.2100 N 131.5400 E

# NCBI SRA aligned data の viewer 表示

The screenshot shows the NCBI SRA search interface. On the left, there are filters for Access (Controlled: 29,708, Public: 78,088), Source (DNA: 1,099, RNA: 100,110), Platform (ABI SOLID: 765, BGISEQ: 57, Complete Genomics: 10), Strategy (other: 107,796), File Type (bam: 54,085, cram: 50,688, fastq: 923, srf: 124), and Other (aligned data: 107,796). A 'Clear all' button is also present. The main search results show items 1 to 20 of 107796, with a summary of 20 per page. The search term used is '(Homo sapiens[Organism]) AND "rna seq"[Strategy]'. The results list includes various Illumina HiSeq 2500 paired end sequencing runs, each with an accession number (e.g., ERX5641105, ERX5641106, ERX5641107, ERX5641108, ERX5641109, ERX5641110) and a download size of 14.5M spots, 2.9G bases, and 366.4Mb.

**①aligned data で絞り込み**

**②Experiment を選び、Run を表示させる**

**③ Alignment を表示**

**④referenceを選択**

**⑤Sequence viewerで表示**

The right side of the screenshot shows the 'Run Browser' for experiment GSM5268398: O\_1h\_A: Optimem 1; Homo sapiens; RNA-Seq (SRR14341180). The 'Alignment' tab is selected. A reference 'chr2' is entered in the 'Reference' input field. The alignment statistics table shows:

Alignment	Reads	Bases	Fraction
Primary	2.7M	130.2Mbp	80.22%

The table below lists the scope of the run:

view	scope	accession	count
	<input checked="" type="radio"/> this run	SRR14341180	1
	<input type="radio"/> same experiment	SRX10695050	1
	<input type="radio"/> same sample	SRS8786418	1
	<input type="radio"/> same study	SRP316699	9
	all sra		33,729

Output options: FASTA format to Screen or File.

# NCBI - Sequence viewer

Homo sapiens chromosome 2, GRCh38.p13 Primary Assembly

gi|568815596|ref|NC\_00002.12|



Run alignments

SRAデータは、パブリッククラウドにも移行されている

- Amazon Web Services (AWS)
- Google Cloud Platform (GCP)

クラウド上で SQL により SRA メタデータ検索を実行できる

- Athena (AWS)
- BigQuery (GCP)

クエリの実行や  
結果の保存は有料

BigQuery

The screenshot shows the BigQuery web interface. The query editor window has the title 'BigQuery' and the sub-title 'FEATURES & INFO'. It displays an 'Unsaved query' with the following SQL code:

```
1 SELECT *
2 FROM `nih-sra-datastore.sra.metadata`
3 WHERE organism = 'Homo sapiens'
```

Below the code, there are several buttons: 'Run', 'Save query', 'Save view', 'Schedule query', and 'More'. A red box highlights a message at the bottom right: 'This query will process 14.7 GB when run.' with a checkmark icon.

<https://www.ncbi.nlm.nih.gov/sra/docs/sra-bigquery-examples/>

# COVID-19

## 新型コロナウィルス関連データポータルサイト

<https://www.ncbi.nlm.nih.gov/sars-cov-2/>

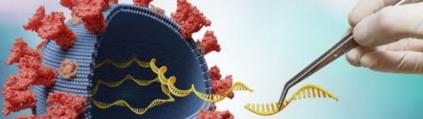
**NCBI SARS-CoV-2 Resources**



**SARS-CoV-2 Data**

664,527	732,293	5,996
SRA runs	Nucleotide records	ClinicalTrials.gov
147,341	166,469	
PubMed	PMC	

**Submit SARS-CoV-2 Sequences**



Add assembled & raw read data to the growing public archive

**SARS-CoV-2 Literature**

Articles referencing SARS-CoV-2 and COVID-19 in PubMed

**View in PubMed**

Free full-text content in PubMed Central (PMC), including preprints, from the Public Health Emergency COVID-19 Initiative, suitable for text mining and secondary analysis

**Access PMC**

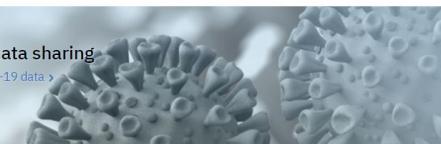
<https://www.covid19dataportal.org/>

**COVID-19 Data Portal**

About ▾ Partners Related resources FAQ Bulk downloads Submit data

Viral Sequences Host Sequences Expression Proteins Biochemistry Imaging Literature

Accelerating research through data sharing  
Read and sign our letter in support of open COVID-19 data ›



**Viral sequences** Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses. 2,158,408 records >

**Host sequences** Raw and assembled sequence and analysis of human and other hosts. 18,733 records >

**VEO variation analysis reports**



**COVID-19 データポータル JAPAN**

概要 連携体制 更新履歴 お問い合わせ en | ja

TOP サイトについて 研究データ 研究ツール European Portal

### データ共有が研究を加速させる

「COVID-19データポータルJAPAN」は、研究に役立つ新型コロナウィルスのデータやサービスを集約し提供します。本システムはEuropean COVID-19 Data Portalの協力の元、JAIRO Cloudの基盤を利用して、国内の多数の機関の援助を受けて公開しています。新型コロナウィルスへの生活への影響や予防については、厚生労働省特設サイトをご覧ください。

**研究データ**  
COVID-19に関する研究データ

遺伝子・配列情報  
タンパク質情報  
リソース情報

**研究ツール**  
日本のCOVID-19研究に役立つサービス

バイオインフォマティクス  
データの投稿

<https://covid19dataportal.jp/>

# NGS データを DRA へ登録する

データファイルの準備

必要なメタデータ情報

メタデータの構成

登録手順

## BAM

- ✓ アライメントされなかったリード(unaligned read)を含めることを推奨
- ✓ リファレンスがある場合、SQ 行のSN 値とリファレンス配列 (INSDC/refseq アクセッション.バージョン番号)との対応表が必要

BAM ファイルヘッダー	対応表
@HD VN:1.0 GO:none SO:coordinate	chr1 NC_000001.11
@SQ SN: <b>chr1</b> LN:249698942	chr2 NC_000002.12
@SQ SN: <b>chr2</b> LN:242508799	chr3 NC_000003.12
@SQ SN: <b>chr3</b> LN:198450956	...
...	

## fastq

- ✓ ペアリードはペアごとに分かれているファイルの登録を推奨
- ✓ gzip か bzip2 で圧縮可能

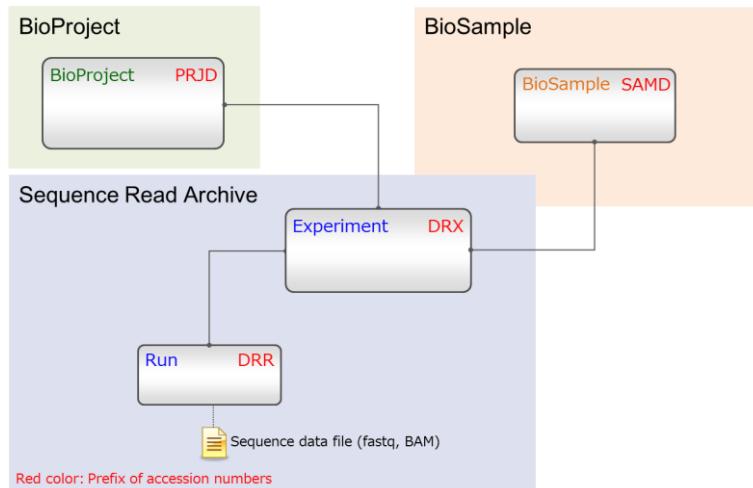
Run データファイル:

<https://www.ddbj.nig.ac.jp/dra/submission.html#run-data-files>

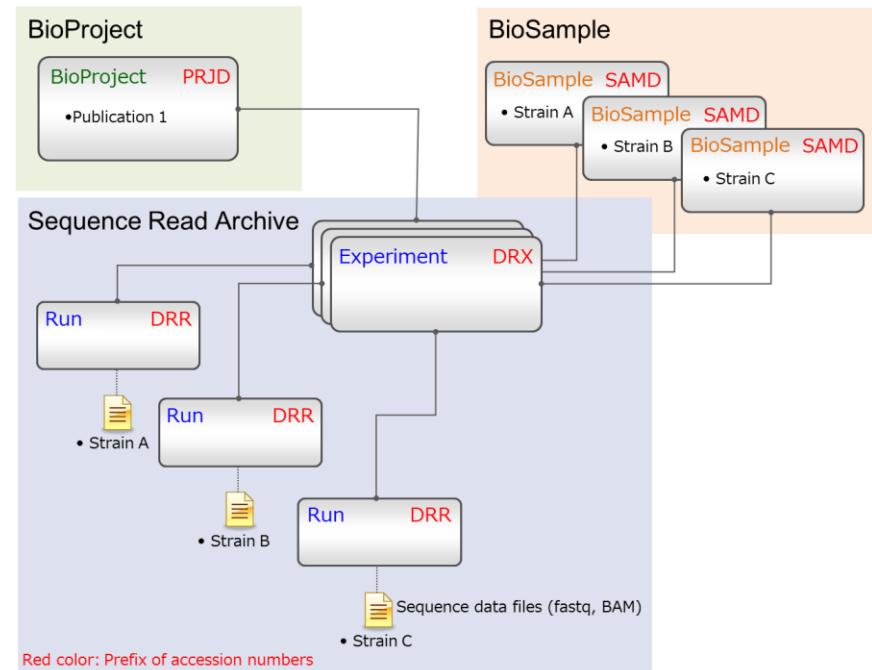
- 登録者情報
  - 登録内容に関して DDBJ からの問い合わせに対応できる方
- 公開予定
  - 即日公開 or 一定期間非公開
    - DRA: 公開予定日を 4 年以内で指定可能
- プロジェクト情報 (BioProject)
  - プロジェクトのタイトル、概要
  - 研究費情報
- 生物学的情報 (BioSample)
  - 生物名 (NCBI Taxonomy database に登録されている種レベル以下の学名)
  - サンプルの採取環境、採取日時、特徴、処理条件など
- シーケンス用ライブラリーとシーケンスの手法 (Experiment)
  - ライブラリの情報
  - シーケンサーの機種名
  - リード長 (spot length)

# DRAメタデータの構成を決めておく

## 1. 最もシンプルな登録



## 2. 三つの菌株の比較ゲノム解析



<https://www.ddbj.nig.ac.jp/dra/submission.html#organization-metadata-objects>

- ✓ 登録する前に必要な BioProject・BioSample・Experiment・Run の数を決めておく
- ✓ サンプル数 ( $\leq$  Experiment/Run 数) から考えると分かりやすい
- ✓ DRA submission では全てのオブジェクトが同時に公開される点に気を付ける

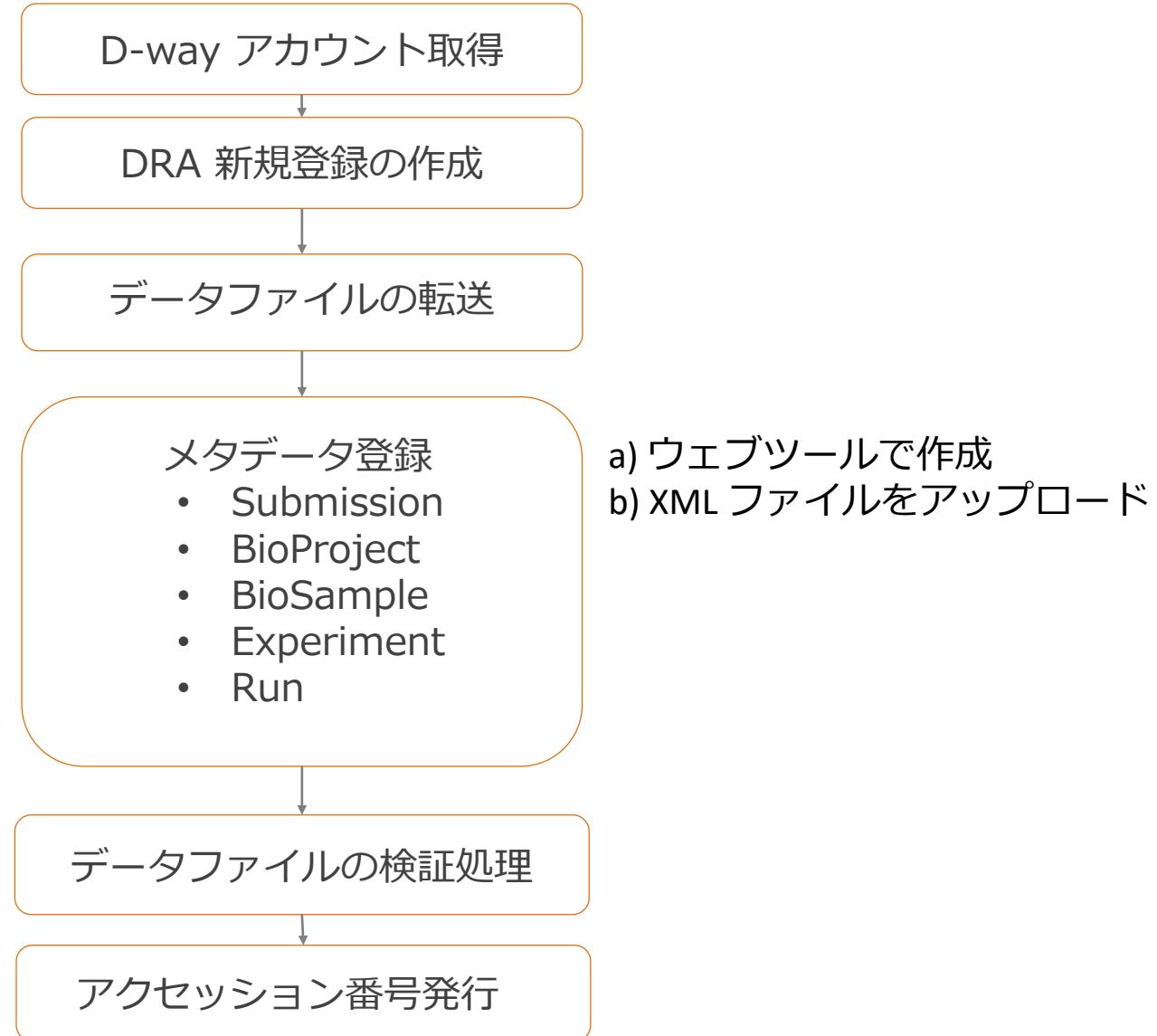
## いくつのサンプルが必要か？

個別の BioSample として扱う場合の例

- ✓ 個体が異なる場合
- ✓ Biological/Technical replicate
- ✓ 組織が異なる場合
- ✓ Time point が異なる場合
- ✓ 処理が異なる場合
  
- 登録例 1 :同じ薬物で処理された3匹の「同一」なトランスジェニックマウス
  - 3 つの Biological replicate = 3 BioSamples
- 登録例 2 :オスのマウス一個体から採取した脳, 心臓, 肺, 精巣, 肝臓
  - 5 つの異なる組織 = 5 BioSamples
- 登録例 3 :ウイルスに感染させた細胞を 0, 2, 4, 8 時間後にサンプリング
  - 4 つの time point = 4 BioSamples

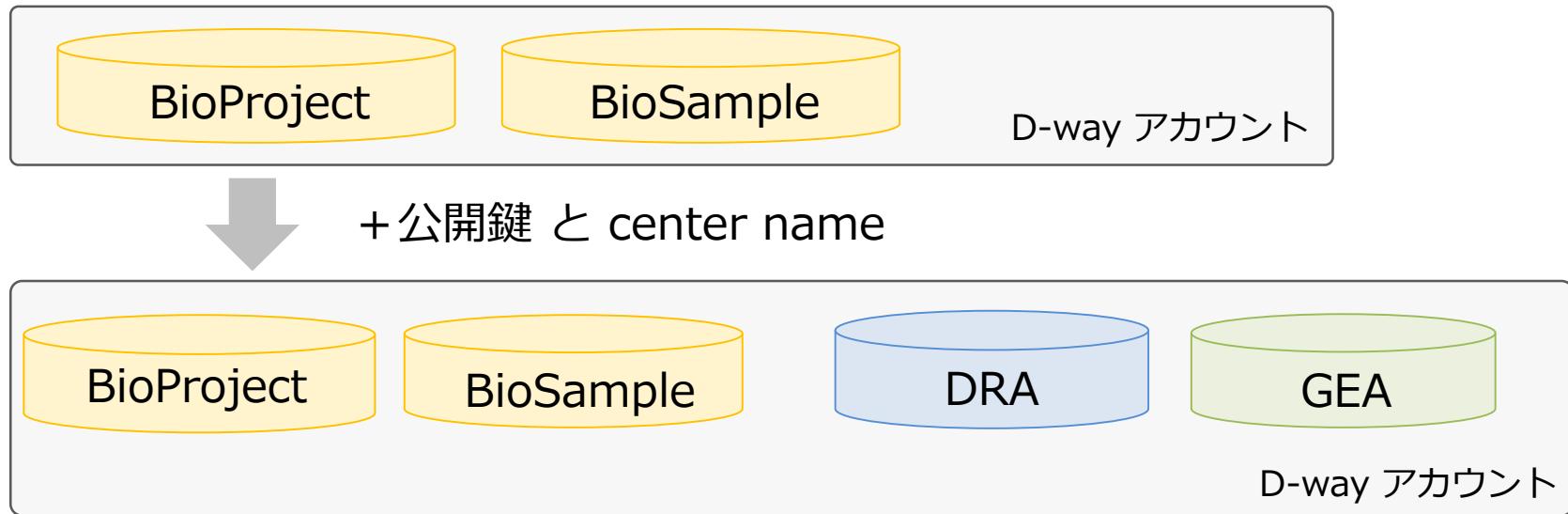
<https://www.ddbj.nig.ac.jp/biosample/submission.html#granularity>

# DRAの登録手順



# D-way アカウントを取得する

- ✓ D-way アカウントをウェブサイト (<https://ddbj.nig.ac.jp/D-way/>) で取得
- ✓ 公開鍵と center name をアカウントに登録し、DRA 登録権限を取得



<https://www.ddbj.nig.ac.jp/account.html>

公開鍵 : 秘密鍵とペアでユーザの認証に使用される  
center name : SRA が組織に運用上割り振っている略号

# DRA Submission の作成

DDBJ DNA Data Bank of Japan D-way TOP | BioProject | BioSample | DRA | Project List | GEA test08 | Account | Password | Logout

## DRA submission list for test08

You can change the hold date of registered DRA submission, update a part of fields (title, library description etc) of DRA metadata from D-way. How to update ([English](#), [Japanese](#)).

New submission Number of submission : 1

DRA Submission ID	DRA Accession	Status	Created Date	Hold Date
<input type="text"/> Reset	<input type="text"/> Reset	<input type="text"/> Reset	2016-08-01	-----
test08-0018		new		

新規作成後、Submission に記載する情報

- ✓ 登録者情報（氏名、メールアドレス、所属組織）
- ✓ 即日公開 or 非公開を選択（**公開予定日を4年以内で指定**）

# データファイルの転送



C:\\$dra		
名前	拡張子	サイズ
..		
test01.fastq	.fastq	41,410 B
test02.fastq	.fastq	41,410 B
test03.fastq	.fastq	41,410 B
test04.fastq	.fastq	41,410 B

/submission/dra/dev/dradev-0019		
名前	拡張子	
..		
test01.fastq	.fastq	



シーケンスデータファイル  
fastq, bam など

DRA ファイル受付サーバ



- ✓ 鍵認証で DRA ファイル受付サーバにアクセスし、データファイルを submission ID に対応するディレクトリ(例: test07-0001) にアップロード

<https://www.ddbj.nig.ac.jp/dra/submission.html#upload-sequence-data>

## DRA Submission ID : test08-0017

DRA Submission ID	DRA Accession	Status	Created Date	Hold Date
test08-0017		new	2016-01-28	-----

**i** All metadata (Experiment, Run, Analysis and referencing BioProject and BioSample) and sequencing data in this submission will be released at the same time.

この Submission に含まれる全てのメタデータ (Experiment, Run, Analysis, 参照されている BioProject と BioSample) とシークエンスデータは同時に公開されます。

1. Upload data files to [ftp-private.ddbj.nig.ac.jp](ftp://ftp-private.ddbj.nig.ac.jp)

データファイルを次のディレクトリにアップ

2. Enter or update metadata in the web interface

ウェブインターフェースでメタデータを新規

[Enter / Update metadata](#)

3. Validate uploaded data files

アップロードしたデータファイルを検証処理にかけてください。

[Validate data files](#)

[Stop validation](#)

\* Stop validation to edit metadata or upload

検証処理を停止してからメタデータを編集もし

[Submit or update metadata by directly uploading](#)

[XML Upload \[+\]](#)

### a) ウェブツールで作成

DRA メニューから BioProject, BioSample を一連の流れでウェブ上で登録できる  
少ない件数の登録向き

### b) XMLファイルをアップロード

エクセルファイルから生成した XML ファイルをアップロード

多数サンプルを一括で登録

Docker, Singularity コンテナを利用

BioProject, BioSample は D-way であらかじめ登録

XML生成方法 <https://github.com/ddbj/submission-excel2xml>

# BioProject の登録

- ✓ D-way 登録画面で各タブに内容を入力して投稿

## SUBMITTER

- ✓ 登録者情報（名前、メールアドレス、所属組織）
- ✓ 即日公開 または 「関連する配列データが公開されるまで非公開」を選択

## GENERAL INFO

- ✓ プロジェクトのタイトル、概要
- ✓ 研究費に関する情報

## PROJECT TYPE

- ✓ プロジェクトのタイプ
  - Genome Sequencing,
  - Transcriptome or Gene Expression, etc.

The screenshot shows the BioProject Submission interface. At the top, there are tabs: SUBMISSION, BIOPROJECT, BIOSAMPLE (highlighted in orange), EXPERIMENT, RUN, and ANALYSIS (optional). Below the tabs, the BioProject Submission ID is displayed as PSUB004192. There are several sub-tabs: SUBMITTER (highlighted in blue), GENERAL INFO, PROJECT TYPE, TARGET, PUBLICATION, and OVERVIEW. The SUBMITTER section contains a table for 'Submitter 1' with three rows: First name (Taro), Last name (Mishima), and E-mail (test08@test.com).

Submitter 1	
First name	Taro
Last name	Mishima
E-mail	test08@test.com

## TARGET

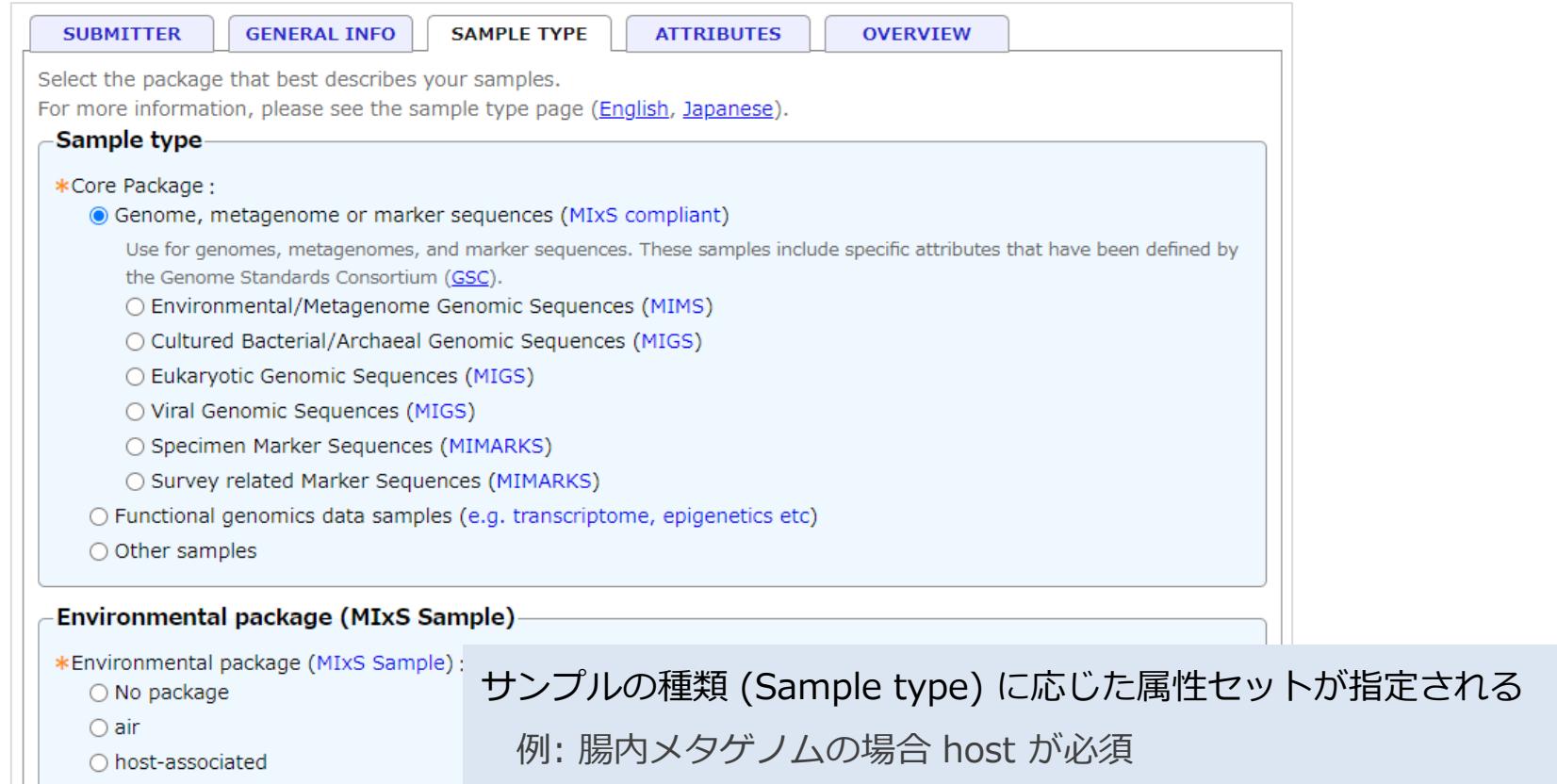
- ✓ 生物情報（対象生物種名）
  - # 複数の生物種を対象としたプロジェクトの場合、共通する階層までの生物分類を記入

## PUBLICATION (任意)

- ✓ 文献情報（Pubmed ID, DOI）
  - # アクセッション番号を記載した論文情報を追加更新

D-way 登録画面で各タブに内容を入力して投稿

- ✓ SUBMITTER: 登録者情報（名前、メールアドレス、所属組織）
- ✓ GENERAL INFO: 即日公開 または 「関連する配列データが公開されるまで非公開」を選択
- ✓ SAMPLE TYPE: サンプルの種類を選択



The screenshot shows the 'SAMPLE TYPE' tab selected in a navigation bar. Below it, a note says: 'Select the package that best describes your samples. For more information, please see the sample type page ([English](#), [Japanese](#)).'. A section titled 'Sample type' contains a required field 'Core Package':

- Genome, metagenome or marker sequences ([MIXS compliant](#))  
Use for genomes, metagenomes, and marker sequences. These samples include specific attributes that have been defined by the Genome Standards Consortium ([GSC](#)).
  - Environmental/Metagenome Genomic Sequences ([MIMS](#))
  - Cultured Bacterial/Archaeal Genomic Sequences ([MIGS](#))
  - Eukaryotic Genomic Sequences ([MIGS](#))
  - Viral Genomic Sequences ([MIGS](#))
  - Specimen Marker Sequences ([MIMARKS](#))
  - Survey related Marker Sequences ([MIMARKS](#))
  - Functional genomics data samples (e.g. transcriptome, epigenetics etc)
  - Other samples

Below this is another section 'Environmental package (MIXS Sample)':

- No package
- air
- host-associated

A callout box highlights the 'Core Package' section with the text: 'サンプルの種類 (Sample type) に応じた属性セットが指定される 例: 腸内メタゲノムの場合 host が必須'.

# BioSampleの登録: ATTRIBUTES

「属性名:値」のペアでサンプルを記述 (例: collection\_date:2015-01-01)

**BioSample Submission ID: SSUB009524**

SUBMITTER GENERAL INFO SAMPLE TYPE ATTRIBUTES OVERVIEW

Describe your sample(s) by providing sample attributes.  
For more information, please see the sample attribute page ([English](#), [Japanese](#))

**Attributes**

\*Attributes file: ファイルを選択 選択されていません

[Download a template text file for BioSample attributes](#)

Download and edit the template text file in spreadsheet or text editor.

- You can submit multiple samples, described in lines.  
サンプルを複数行に記載し、まとめて登録することができます。
- If you do not have information for a required attribute(s), please provide the value as either 'not collected', 'not applicable' or 'missing'.  
必須属性に対する値がない場合は、'not collected', 'not applicable' もしくは 'missing' を記入してください。
- Provide 'organism' and 'taxonomy\_id' as indicated in [NCBI Taxonomy \(for unregistered organism\)](#).  
[NCBI Taxonomy](#)に登録されている生物を、「organism」と'taxonomy\_id'に記入してください (未登録の生物の場合)

Continue

*sample_name	*sample_title	description	*organism	taxonomy_id
spotX	soil metagenome in spot X		soil metagenome	410658
spotY	soil metagenome in spot Y		soil metagenome	410658

- ✓ Sample type に対応したタブ区切りのテキストファイルをダウンロード
- ✓ エクセルなどで 1 行に 1 サンプルの情報を入力し、テキストファイルをアップロード
- ✓ 複数サンプルの登録が可能

# サンプル属性一覧

DDBJ サービス スパコン 統計 活動 センターについて DDBJ Web Sites ▾  利用規約 問合せ

## BioSample

Home Handbook Sample Attribute ▾ Validation Rules FAQ Search Downloads ▾ About BioSample

ホーム > biosample > サンプル属性

### サンプル属性

List all sample attributes

Sample type (Core Package)

- Genome, metagenome or marker sequences (MiXs compliant)
- Functional genomics samples (e.g. transcriptome, epigenetics etc.)
- Other samples (e.g. transcriptome, epigenetics etc.)

**DEFINITION DOWNLOAD**

Sample type を選択し、DEFINITION ボタンで attribute の定義と書式を見ることができます。DOWNLOAD ボタンで BioSample ワークシートをダウンロードすることができます。定義表  いくつかのパッケージの登録例  を公開しています。

データ種別毎のサンプル登録

- [Human Sample](#)
- [Metagenome Assembly](#)
- [Single amplified genome](#)
- [Pseudohaplotype](#)

All attributes

Example: All 

Name	Description
<a href="#">sample_name</a>	sample name (は登録者がサンプルに付ける名前です。sample name は Submission においてユニークである必要があります。最大100文字で英数字、半角空白と ()[]+_- から構成されている必要があります)。
<a href="#">sample_title</a>	タイトルはサンプルをよく表す簡潔なものを記入します。タイトルは Submission においてユニークである必要があります。例: 1) Escherichia coli O104:H4 str. C227-11 clinical isolate 2010_333_NC-6; 2) CD8+ T cells from female TSG6-knockout BALB/c mouse; 3) Human metagenome isolated from urine of healthy female.
<a href="#">description</a>	サンプルに対する簡潔な補足情報。
<a href="#">organism</a>	NCBI Taxonomy database  に登録されている最も下位のランクの生物名 (適切な場合は species まで)。データベースに登録されていない場合、未登録の生物に関する情報をできるだけ記入してください。DDBJ スタッフが NCBI Taxonomy に未登録の生物を申請します。



サンプル属性一覧: <https://www.ddbj.nig.ac.jp/biosample/attribute.html?all=all>

# BioSample 登録例

BioSample Examples ☆ ☁ 🔍

ファイル 編集 表示挿入 表示形式 データ ツール アドオン ヘルプ

🖨️ 🔍 100% 🌐 閲覧のみ

B1 | fx |

	A	B	
1	Functional.genomics		
2	*sample_name	*sample_title	description
3	Arabidopsis control for heat stress rep 1	Arabidopsis control for heat stress rep 1	Arabidopsis contr
4	Arabidopsis control for heat stress rep 2	Arabidopsis control for heat stress rep 2	Arabidopsis contr
5	Arabidopsis heat stress rep 1	Arabidopsis heat stress rep 1	Arabidopsis heat
6	Arabidopsis heat stress rep 2	Arabidopsis heat stress rep 2	Arabidopsis heat
7			
8	MIMS.me.host-associated		
9	*sample_name	*sample_title	description
10	L1_1	gut content sample of silver carp no.1_1	
11	L1_2	gut content sample of silver carp no.1_2	
12	L2_1	gut content sample of silver carp no.2_1	
13	L2_2	gut content sample of silver carp no.2_2	
14	LB_1	gut mucosa sample of silver carp 1	
15	LB_2	gut mucosa sample of silver carp 2	
16	W1_1	gut content sample of Wuchang bream no.1_1	
17	W1_2	gut content sample of Wuchang bream no.1_2	
18	W2_1	gut content sample of Wuchang bream no.2_1	
19	W2_2	gut content sample of Wuchang bream no.2_2	
20	WB_1	gut mucosa sample of Wuchang bream 1	
21	WB_2	gut mucosa sample of Wuchang bream 2	
22			
23	MIMS.me.human-gut		
24	*sample_name	*sample_title	description
25	patient 1	gut microbiomes isolated from feces of Japanese patient 1	
26	patient 2	gut microbiomes isolated from feces of Japanese patient 2	
27	patient 3	gut microbiomes isolated from feces of Japanese patient 3	
28	patient 4	gut microbiomes isolated from feces of Japanese patient 4	
29			
30	MIMS.me.miscellaneous		

All ▾ Functional.genomics ▾ MIMS.me.host-associated ▾ MIMS.me.human-gut ▾ MIMS.me.miscellaneous ▾

# DRA Experiment, Run の作成

## Experiment

- ✓ サンプルから構築したライブラリー、シークエンサーやリード長について記入

記入例	
BioSample Used	SAMD00000001
Library Source	GENOMIC
Library Selection	RANDOM
Library Strategy	WGS
Instrument	Sequel
Spot Type	Single
Spot Length	30000 (不定長の場合はリード長の平均)

## Run

- ✓ 対応する Experiment 番号
- ✓ データファイルの MD5, file type を記述
- ✓ 不定長の fastq の場合、file type は generic\_fastq を選択

# データファイルの検証処理

DRA Submission ID : test08-0017

DRA Submission ID	DRA Accession	Status	Created Date	Hold Date
test08-0017		metadata_submitted	2016-01-28	2018-01-28

1. Upload data files to dradata.ddbj.nig.ac.jp:/submission/test08/test08-0017

2. Enter or update metadata in the web interface.

Enter / Update metadata

3. Validate uploaded data files

Validate data files

Stop validation

\* Stop validation to edit metadata or upload data files

Upload and validate this submission

検証処理を開始

with the submitted metadata to start reviewing process.  
the status become "submission\_validated" or "data\_error"

- ✓ データファイルの形式とメタデータとの整合性が検証され、アーカイブ用の SRA ファイルが作成される

# アクセッショント番号の発行

DRA Submission ID : test07-0018

DRA Submission ID	Accession #	Status	Creation Date	Hold Date
test07-0018	DRA003024	completed	2015-03-09	2017-03-09 Change

Submit / Update Metadata

Validate data files Stop validation

\* Stop validation to edit metadata or upload data files

XML Upload [+]

Component [-]

Object	BioProject ID	BioSample ID	Accession #	Center Name	Alias
submission			DRA003024	DDBJ	test07-0018_Submission
+ experiment	PRJDB3521	SAMD00025505	DRX026751	DDBJ	test07-0018_Experiment_0001
+ run			DRR029693	DDBJ	test07-0018_Run_0004
+ experiment	PRJDB3521	SAMD00025506	DRX026752	DDBJ	test07-0018_Experiment_0002
+ run			DRR029694	DDBJ	test07-0018_Run_0005
+ experiment	PRJDB3521	SAMD00025507	DRX026753	DDBJ	test07-0018_Experiment_0003
+ run			DRR029695	DDBJ	test07-0018_Run_0006

- ✓ BioProject (PRJD)
- ✓ BioSample (SAMD)
- ✓ Submission (DRA), Experiment (DRX), Run (DRR)

論文には「データ登録」に対する DRA アクセッショント番号の引用を推奨

# データの公開

- ✓ 公開されたデータはミラーされ DDBJ/EBI/NCBI で利用できる

## DRASearch

The screenshot shows the DRASearch interface. On the left, there's a 'Submission Detail' section with the following information:

Alias	DRA000001
Submission ID	
Submission Date	2009-05-14
Center Name	KEIO
Lab Name	Bioinformatics Lab.

On the right, there's a 'Navigation' section with links to 'Study DRP000001', 'Experiment DRX000001', 'Sample DRS000001', and 'Run DRR000001'. Each link has download icons for FASTQ and SRA.

## ENA

The screenshot shows the ENA interface for run DRR000001. It displays sequencing parameters and a table of sample details.

**Sequencing Parameters:**

- Organism: *Bacillus subtilis* subsp. *natto* BEST195
- Sample Accession: SAMD00016353
- Instrument Platform: ILLUMINA
- Instrument Model: Illumina Genome Analyzer II
- Read Count: 10148174
- Base Count: 73066528
- Center Name: KEIO
- Library Layout: PAIRED
- Library Strategy: WGS
- Library Source: GENOMIC

**Read Files:**

Download report: JSON TSV

Download Files as ZIP Download Selected files

Study Accession	Sample Accession	Experiment Accession	Run Accession	Tax Id	Scientific Name	FASTQ FTP	Submitter
PRJDA38027	SAMD00016353	DRX000001	DRR000001	645657	<i>Bacillus subtilis</i> subsp. <i>natto</i> BEST195	<input type="checkbox"/> DRR000001_1.fastq.gz <input type="checkbox"/> DRR000001_2.fastq.gz	

## NCBI SRA

The screenshot shows the NCBI SRA interface for the same sequencing run. It lists individual reads with their names and coordinates.

**Reads (separated):**

- DRR000001.1 DRS000001
- DRR000001.2 DRS000001
- DRR000001.3 DRS000001
- DRR000001.4 DRS000001
- DRR000001.5 DRS000001
- DRR000001.6 DRS000001
- DRR000001.7 DRS000001

Sequence data for the first few reads is shown:

```
>gnl|SRA|DRR000001.1.3060N:7:1:1116:340 (Biological)
GATGGTAAGATAAGCAGTTGAAGTTACAAACCG
>gnl|SRA|DRR000001.2.3060N:7:1:1116:340 (Biological)
GACACGTCCCTTTCATCATTCACCTTGACTTGAAT
```