統合データベース講習会：AJACSオンライン7
2021年7月15日

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

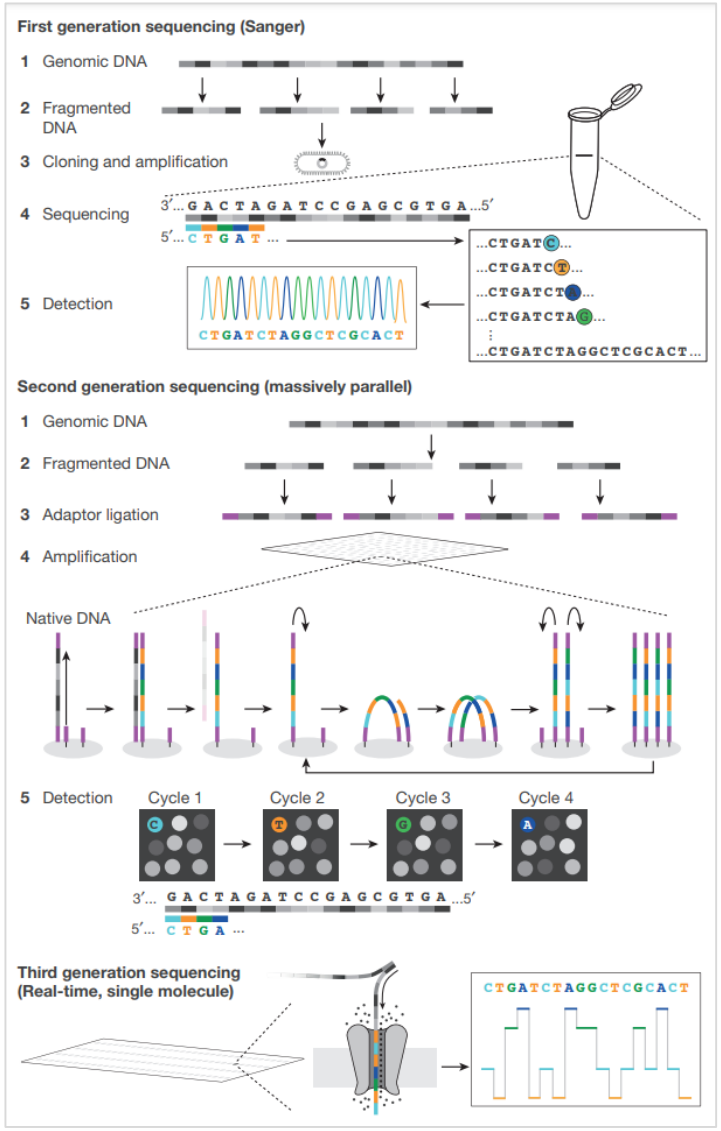
福田　亜沙美
国立遺伝学研究所 生命情報・DDBJセンター
キュレータ

講義の内容

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)

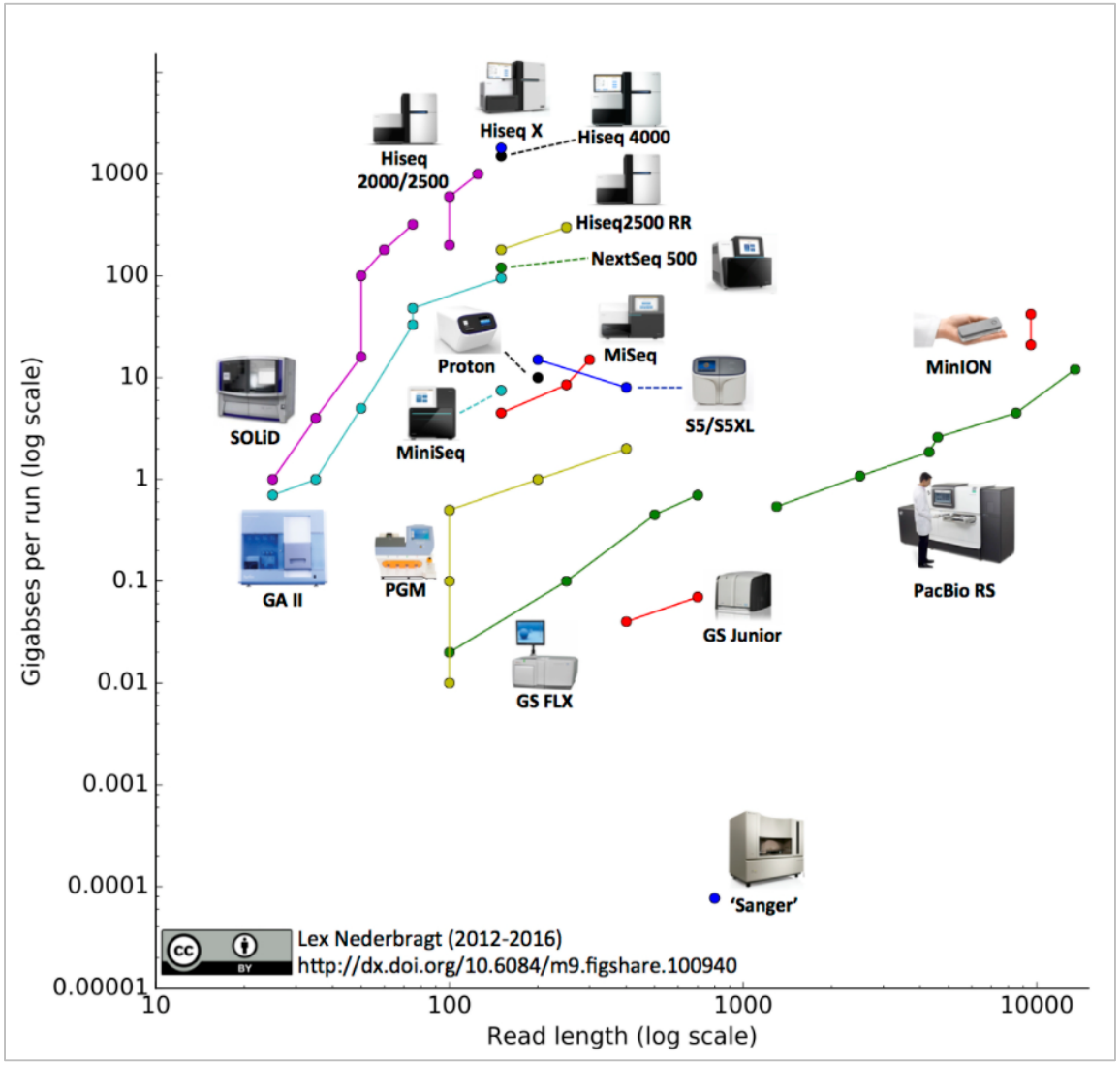
これ以降を次世代シーケンサーと呼ぶ

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



・Shor read はさらに大量出力

Illunima 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…
+
IIIIIII9IG9IC7…
```

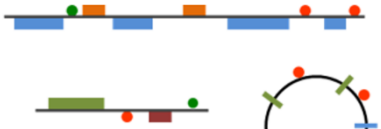
アライメント

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

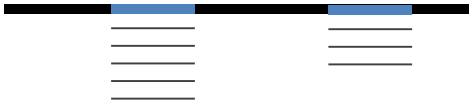


アノテーション

fasta

```
>seq1
GGGTGATGGCCGC
//
```

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.

	s1	s2	s3
ENSMUSG00000109643	1	1	1
ENSMUSG00000109644	0	0	1
ENSMUSG00000109645	0	1	1

バリアントデータ

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

NGS 生データ

Sequence Read Archive

DRA
ENA
SRA

fastq

```
@seq1
GGGTGATGGCCGCT…
+
IIIIIII9IG9IC7…
```

bam

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

DDBJ Center
EBI
NCBI

アセンブル/アノテーション付き配列データ

DDBJ
ENA
GenBank

ENA: European Nucleotide Archive

機能ゲノミクスデータ

GEA
ArrayExpress
GEO

※ArrayExpress→BioStudiesに移行予定

GEA:Genomic Expression Archive
GEO:Gene Expression Omnibus

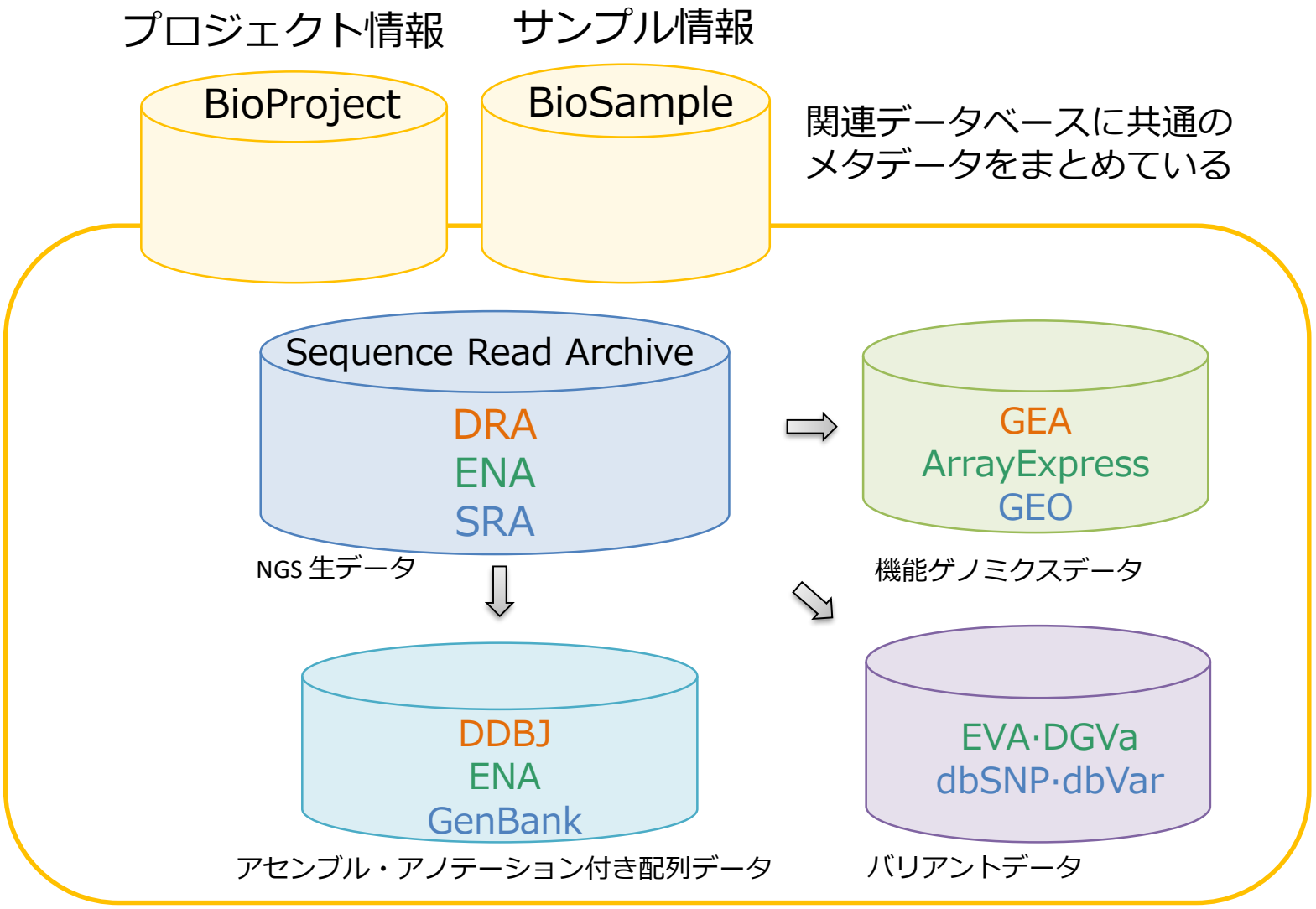
バリアントデータ

EVA·DGVa
dbSNP·dbVar

EVA・DGVa 全生物種
dbSNP・dbVar ヒトデータ限定

EVA:European Variation Archive
DGVa:The Database of Genomic Variants archive

データを整理するデータベース

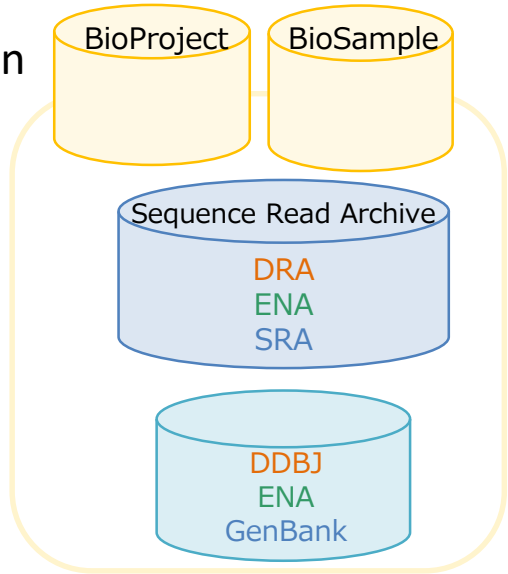


※メタデータ: シークエンスデータを説明するデータ

INSDCデータベース

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	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

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

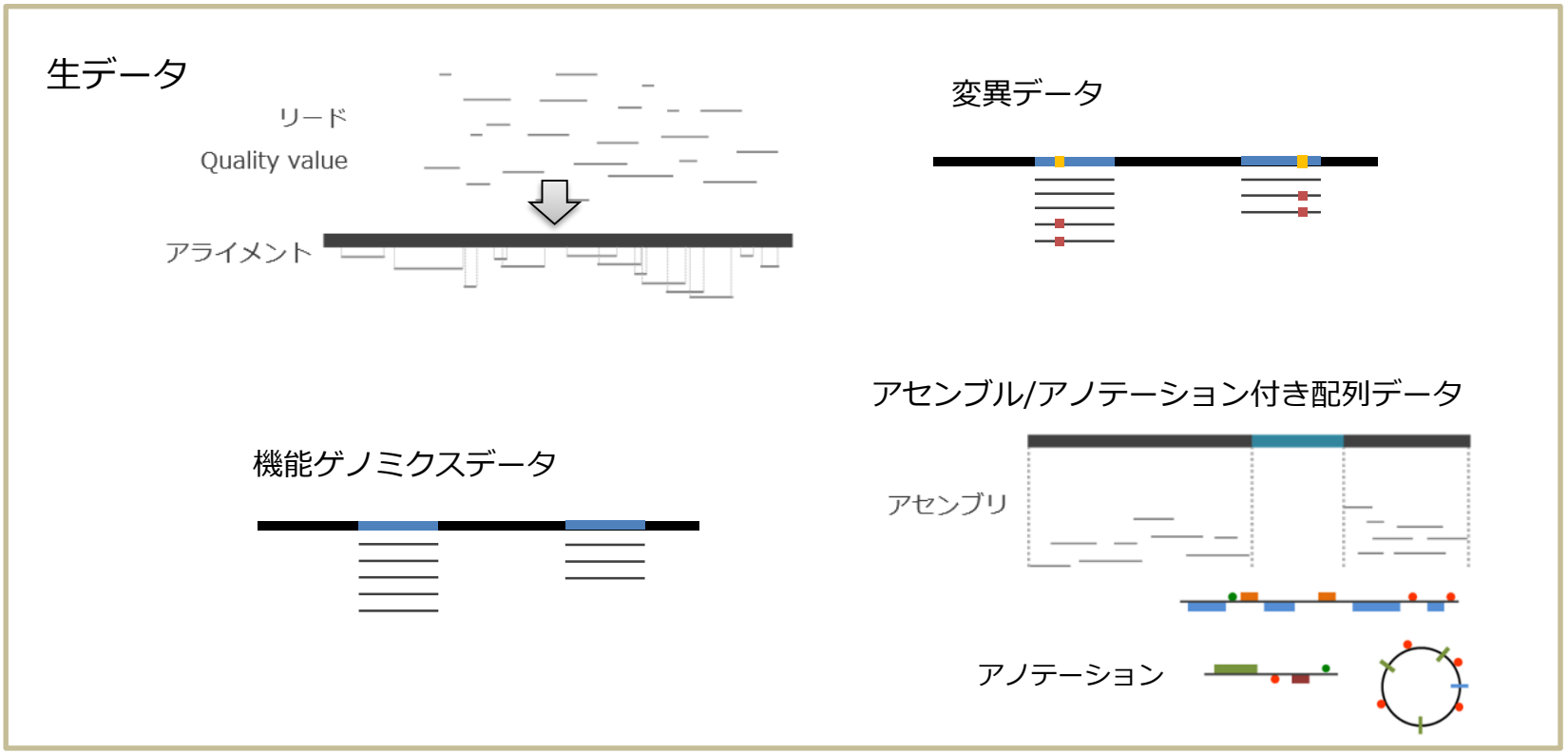
↓ アクセス制限

↓ オープンアクセス

Japanese Genotype-phenotype Archive (JGA) ※NBDCと共同運営

European Genome-Phenome Archive (EGA)

The database of Genotypes and Phenotypes (dbGAP)



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	•DRA/ERA/SRA (INSDC) •DDBJ/ENA/GenBank (INSDC) •GEA/ArrayExpress/GEO •EVA·DGVa/dbSNP·dbVar •PDB	•RefSeq •Ensembl •Expression Atlas •ChIP-Atlas •UniProt

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

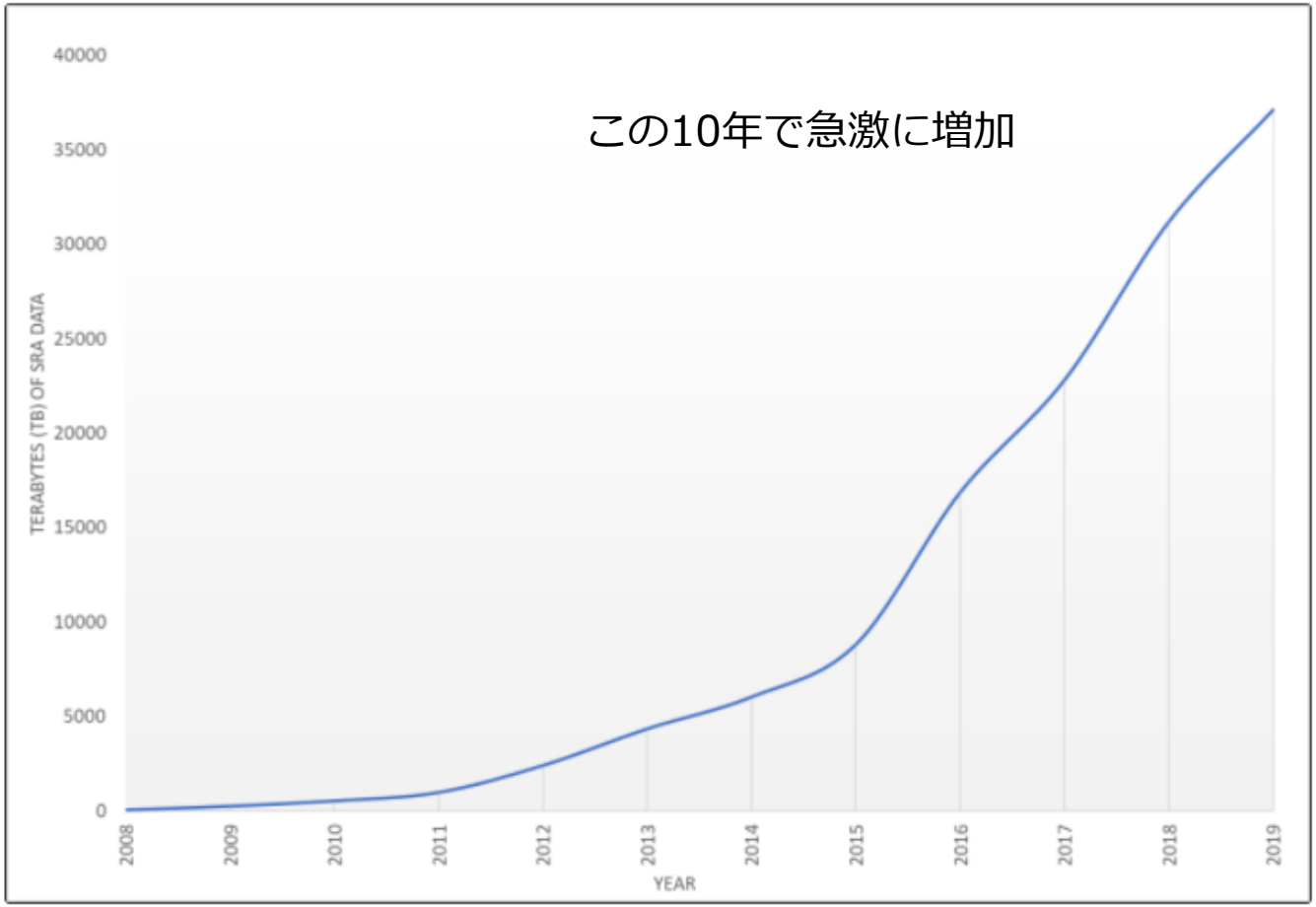


Figure 1. SRA data has grown exponentially over the last decade.

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…
+
IIIIIII9IG9IC7…
```

bam

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

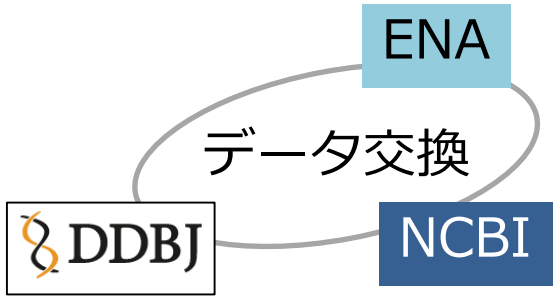
シーケンサー出力
(PacBio hdf5 etc.)

SRA Toolkit →

アーカイブ用 SRA ファイル

SRA Toolkit →

fastq

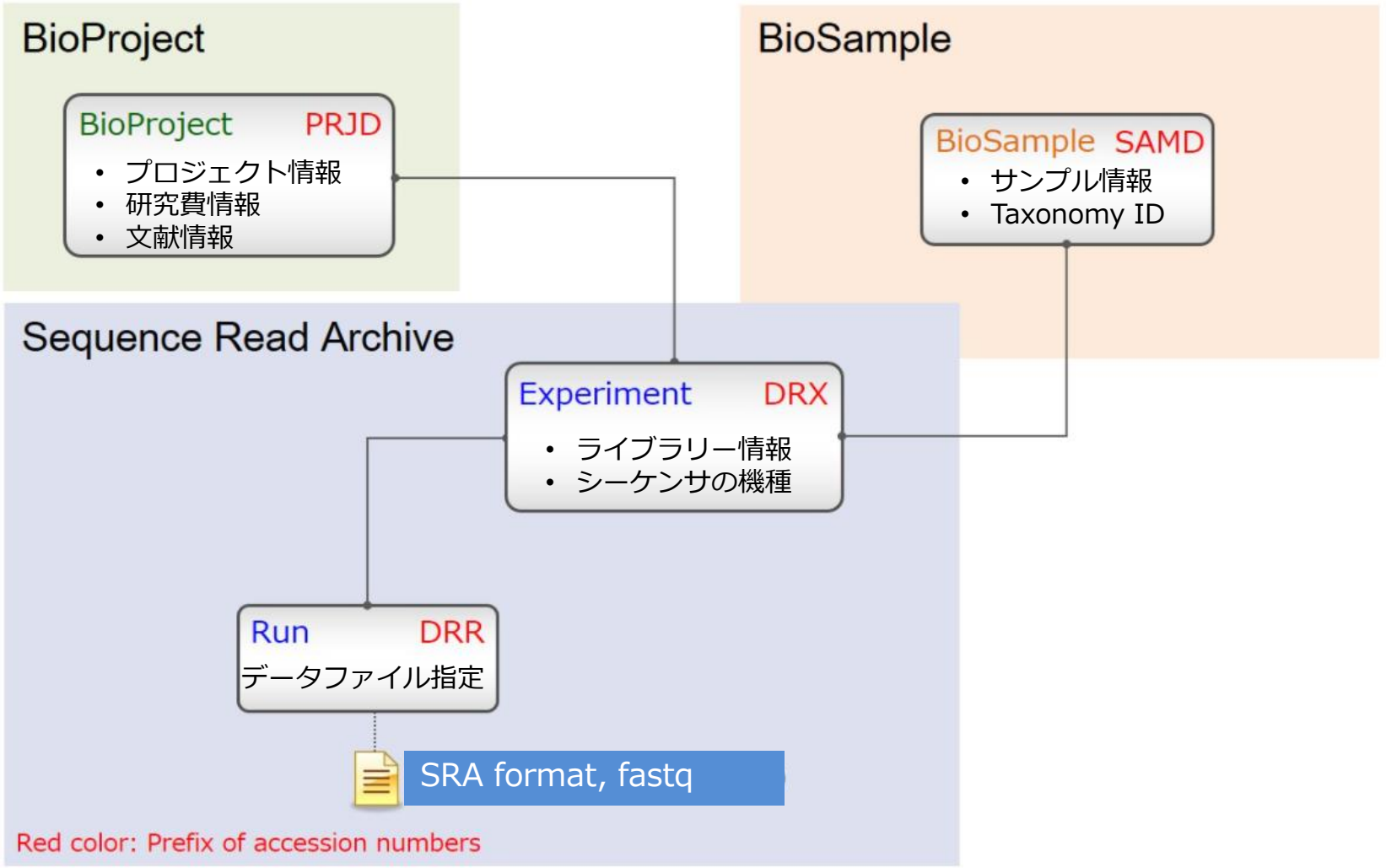


公開データフォーマット

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

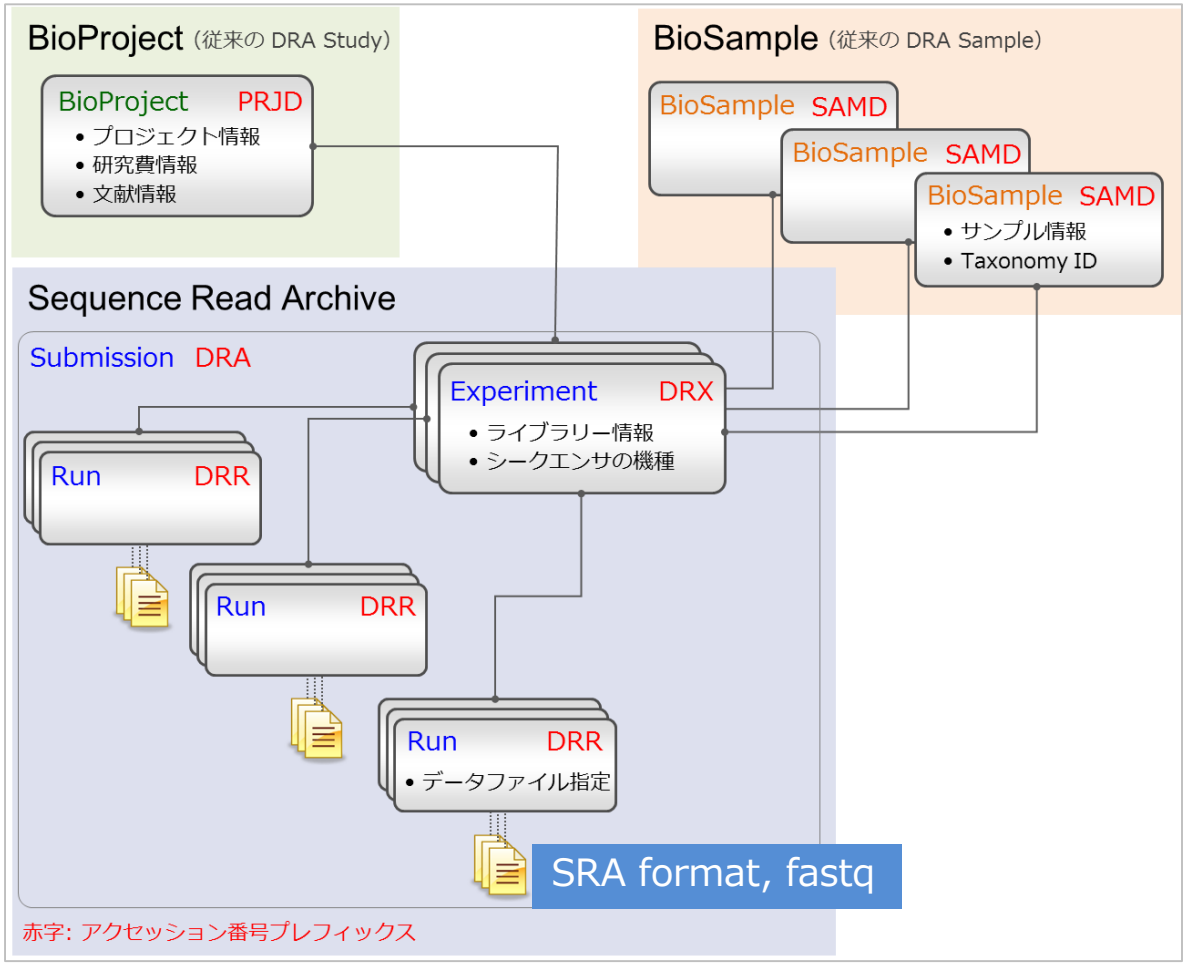
SRAメタデータ



✓ SRA データは３つのデータベースにまたがる

SRAメタデータ

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



BioProject

プロジェクト概要

生物情報

論文情報

関連データ

SRA データ

BioSample

Assembly

SRA
データサイズ

Perna viridis (Asian green mussel)

Accession: PRJDB10567 ID: 7

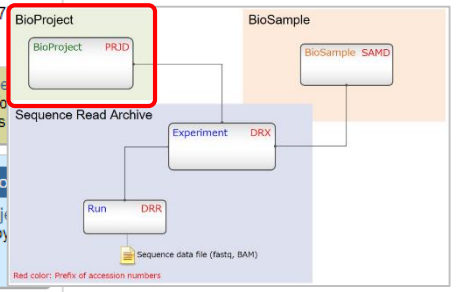
Asian green mussel whole genome sequencing

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

See Genome Information fo Perna viridis

NAVIGATE ACRO

8 additional proje are related by organism.



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

Entoria okinaw
RNA sequencing

アクセッション番号プレフィックス
- PRJD (DDBJ)
- PRJE (ENA)
- PRJN (NCBI)

Project Data:

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

- ✓ 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

▾ Assembly details: Download

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

NCBI BioProject:https://www.ncbi.nlm.nih.gov/bioproject/PRJDB10567

BioSample

タイトル → **Wild Perna viridis from Enoshima**

Identifiers	BioSample: SAMD00247165; SRA: DRS176506
生物情報 → Organism	Perna viridis (Asian green mussel) cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Spiralia; Lophotrochozoa; Mollusca; Bival Pteriomorpha; Mytiloida; Mytiloidea; Mytilidae; Mytilinae; Perna
パッケージ → Package	MIGS: eukaryote; version 5.0

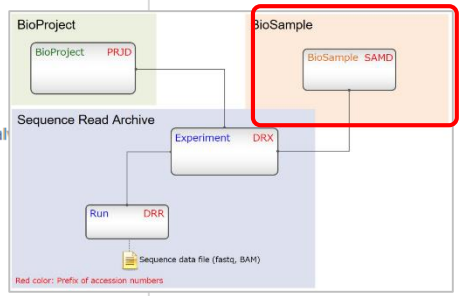
サンプル属性 → Attributes

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	Japan:Kanagawa, Enoshima
isolation and growth condition	https://www.jstage.jst.go.jp/article/sosj1996/21/1/21_1_19/_article/-char/ja/
latitude and longitude	35.18 N 139.28 E
number of replicons	missing
ploidy	diploid
project name	Asian green mussel whole genome sequen
propagation	sexual
cultivar	missing
ecotype	missing
isolate	missing
strain	missing

Description	Keywords: GSC:MIxS;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

関連データ → BioProject SRA Nucleotide



アクセッ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

SRA Experiment

DDBJ DNA Data Bank of Japan

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

Submitted by: AORI

BioProject → **Study:** Asian green mussel whole genome sequencing
PRJDB10567 • DRP007082 • All experiments • All runs
show Abstract

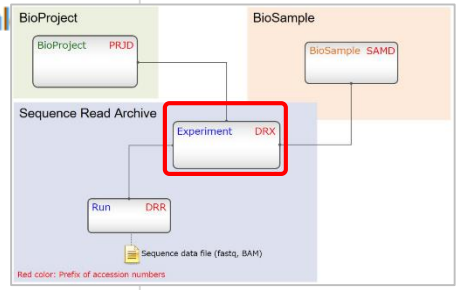
BioSample → **Sample:** Wild Perna viridis from Enoshima
SAMD00247165 • DRS176506 • All experiments • All runs
Organism: Perna viridis

ライブラリー →
シークエンサー → **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
DRR244978	39,412,069	7.9G	2.9Gb	2021-03-25



アクセッ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

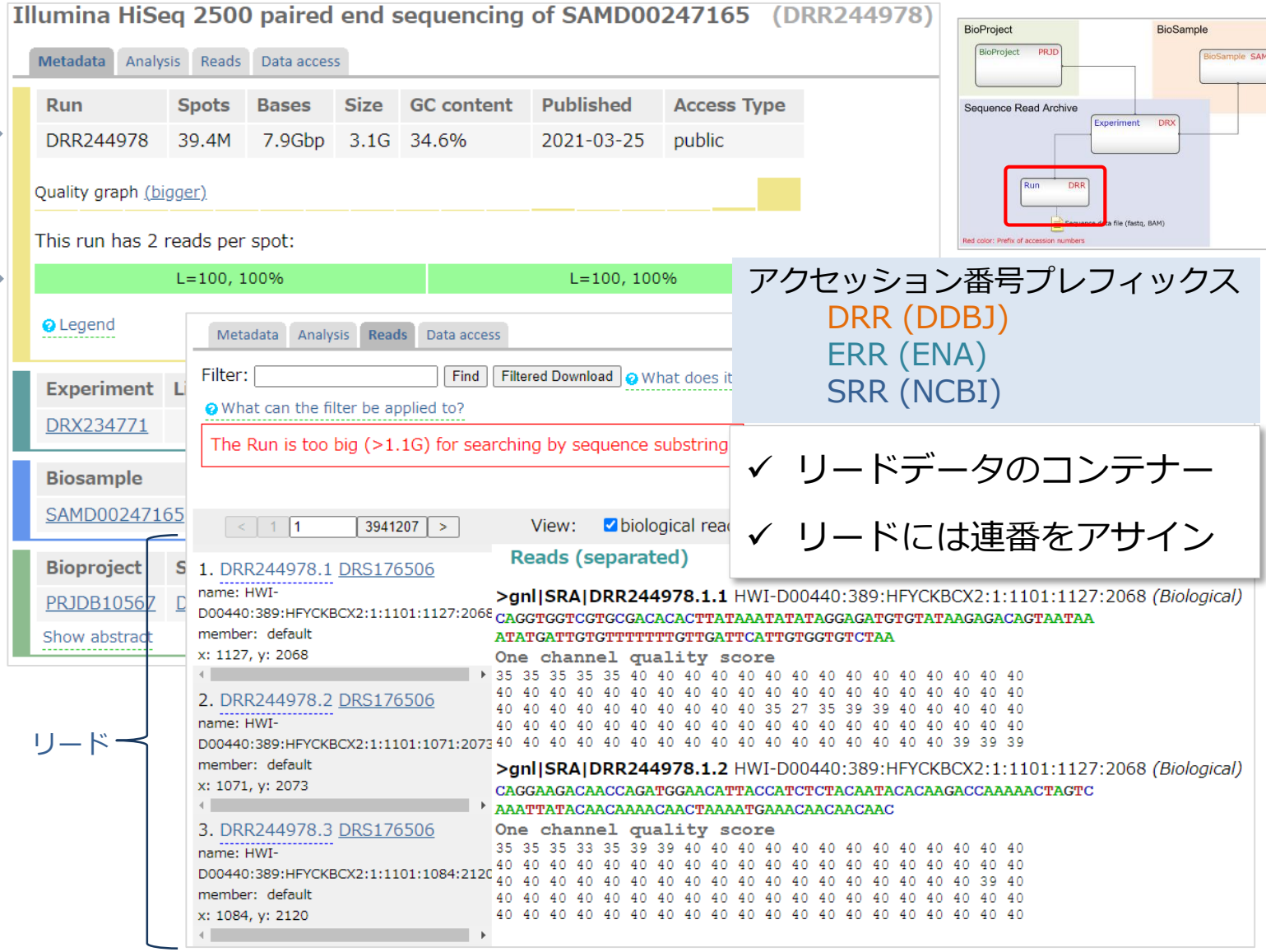
NCBI Experiment: https://www.ncbi.nlm.nih.gov/sra/DRX234771

SRA Run

Run 情報

リード構成

リード



アクセッ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

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 (*)

- Advanced Search

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

https://www.ebi.ac.uk/ebisearch/documentation.ebi#query_syntax

DRASearch

https://ddbj.nig.ac.jp/DRASearch/

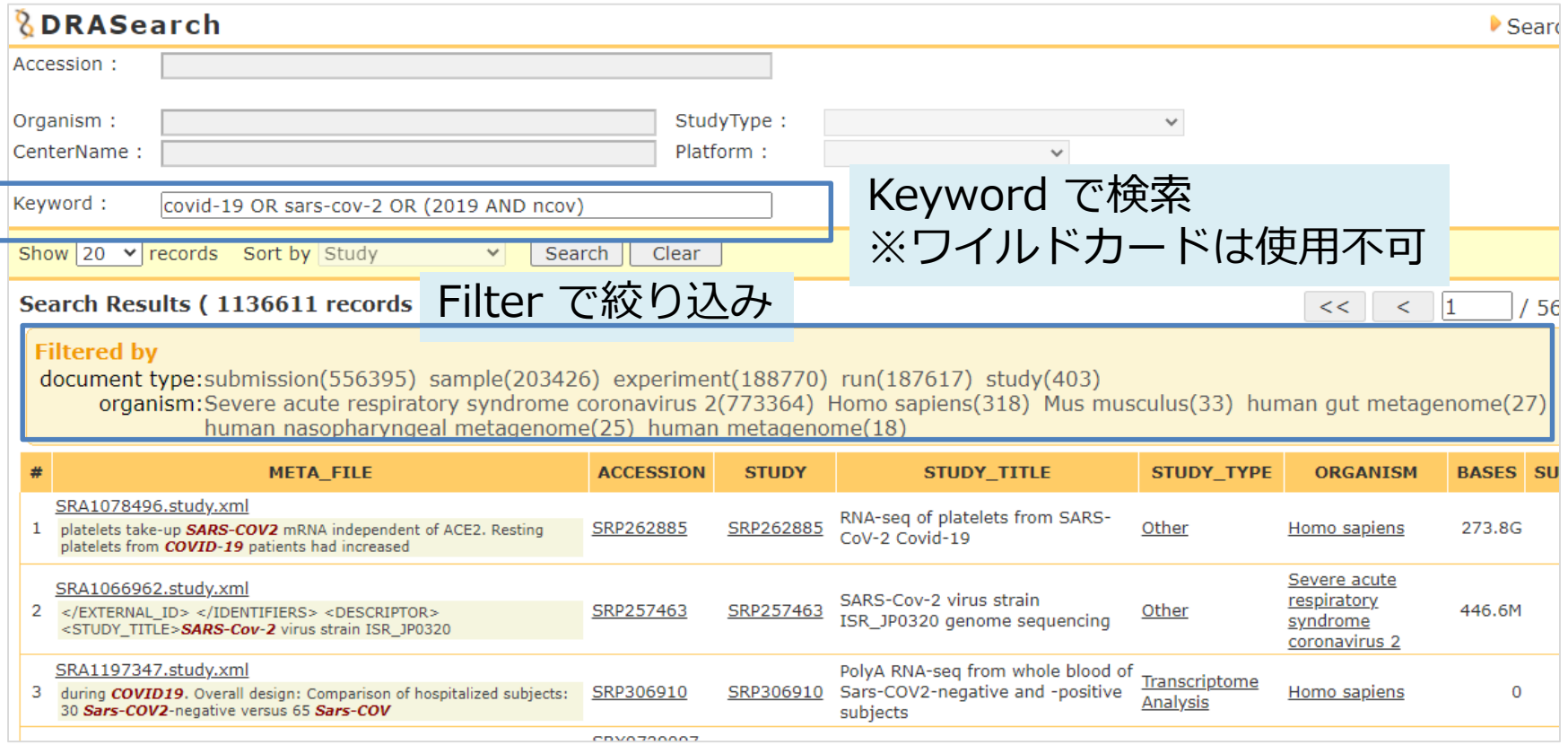
DDBJ DNA Data Bank of Japan

対象データは DRA/ERA/SRA

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

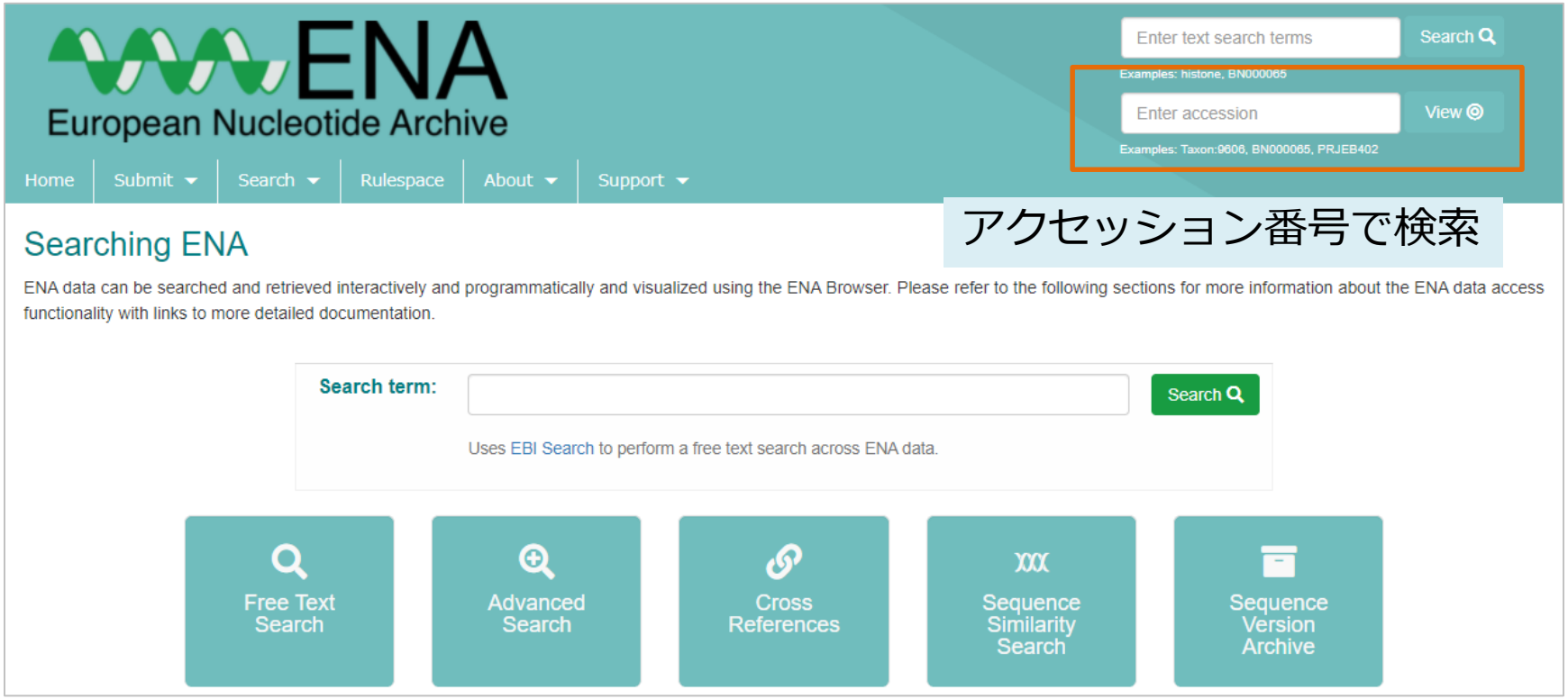


DBCLS SRA

https://sra.dbcls.jp/

ENA browser

https://www.ebi.ac.uk/ena/browser/search



アクセッション番号で検索

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	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

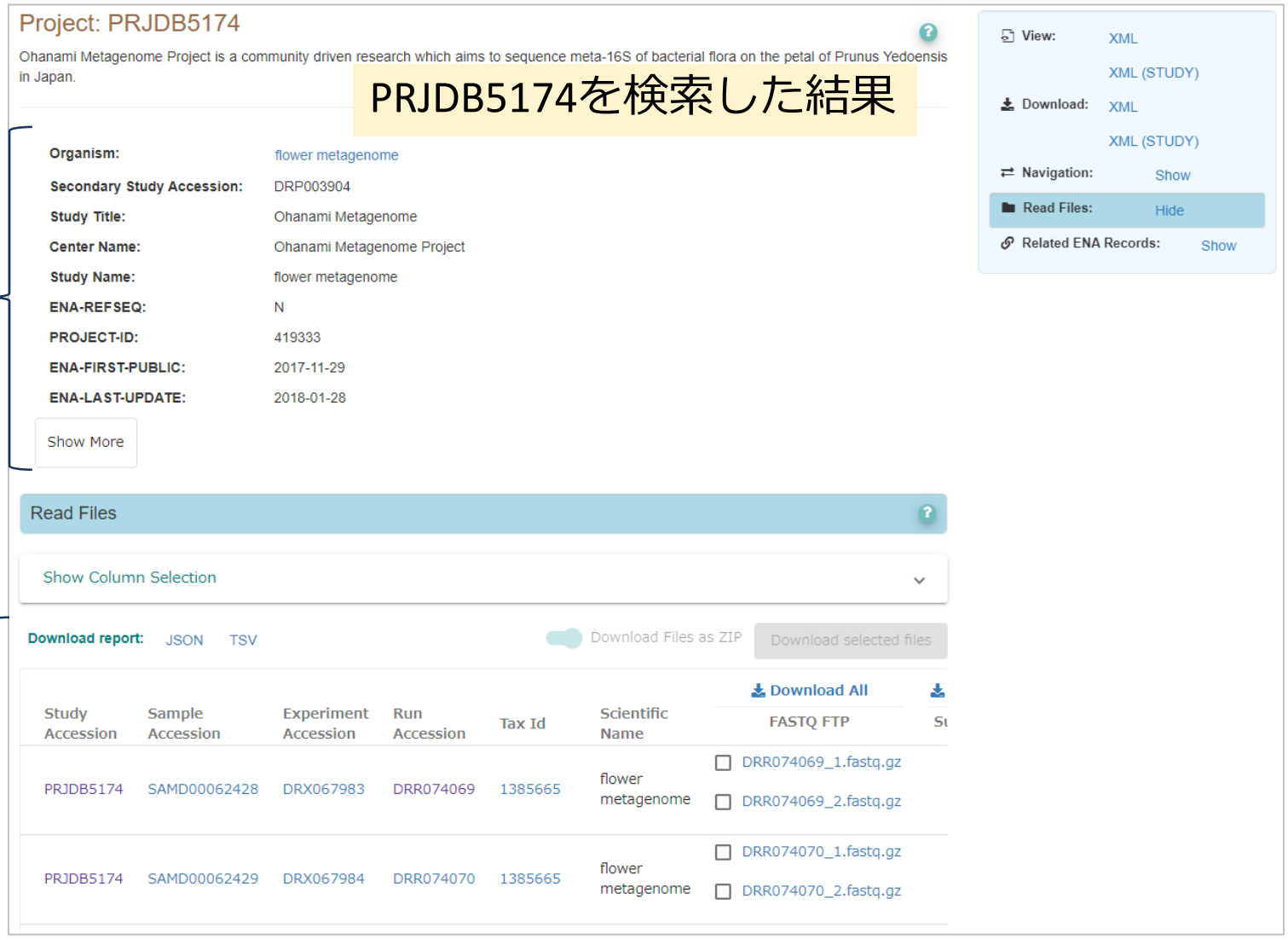
2020年8月から新ブラウザへ完全移行
https://www.ebi.ac.uk/about/news/service-news/retirement-old-ena-browser-5th-august-2020

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



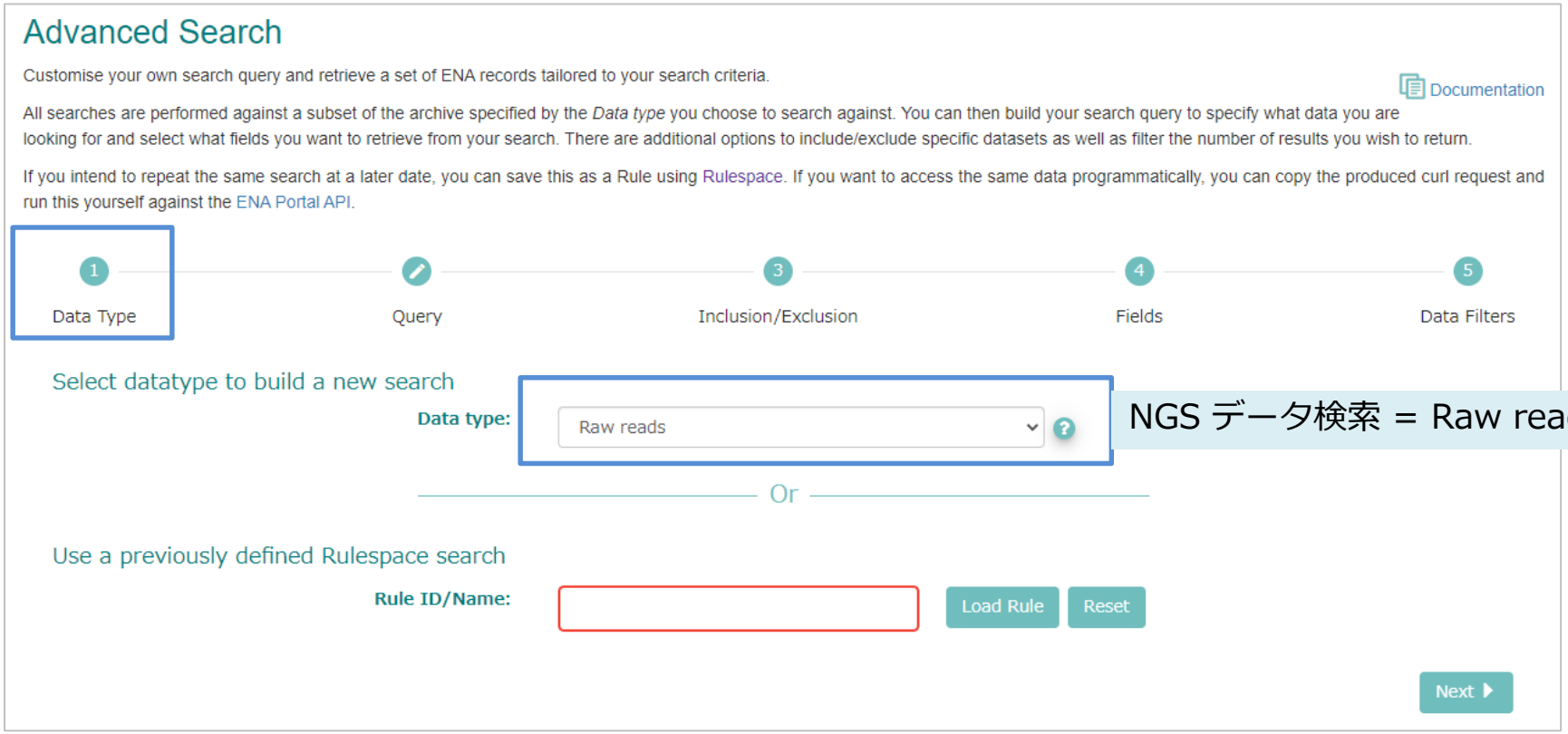
https://www.ebi.ac.uk/ena/browser/view/PRJDB5174

ENA browser – keyword 検索

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



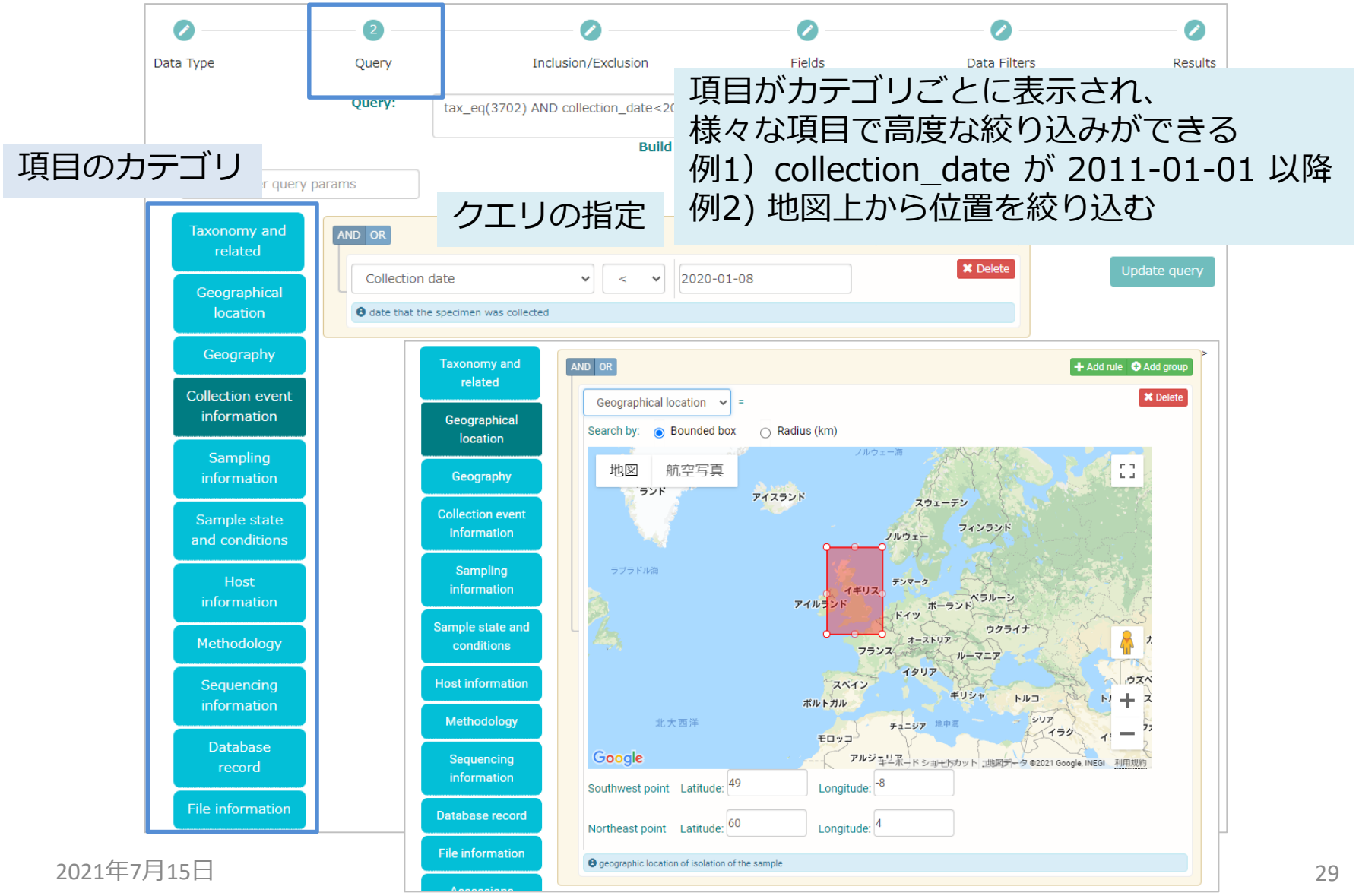
NGS データ検索 = Raw reads

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

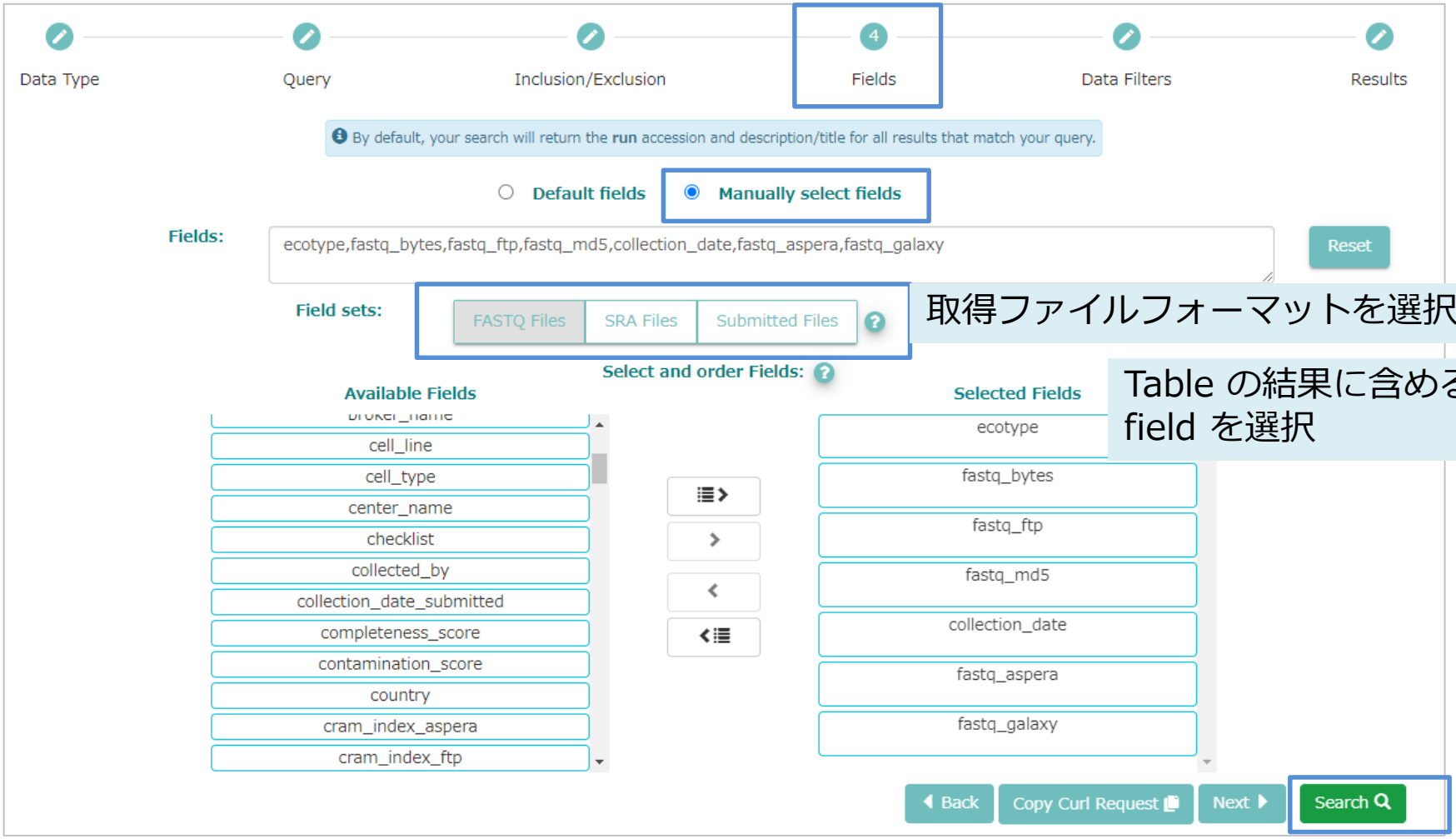
Advanced Search - Query

クエリを細かく指定



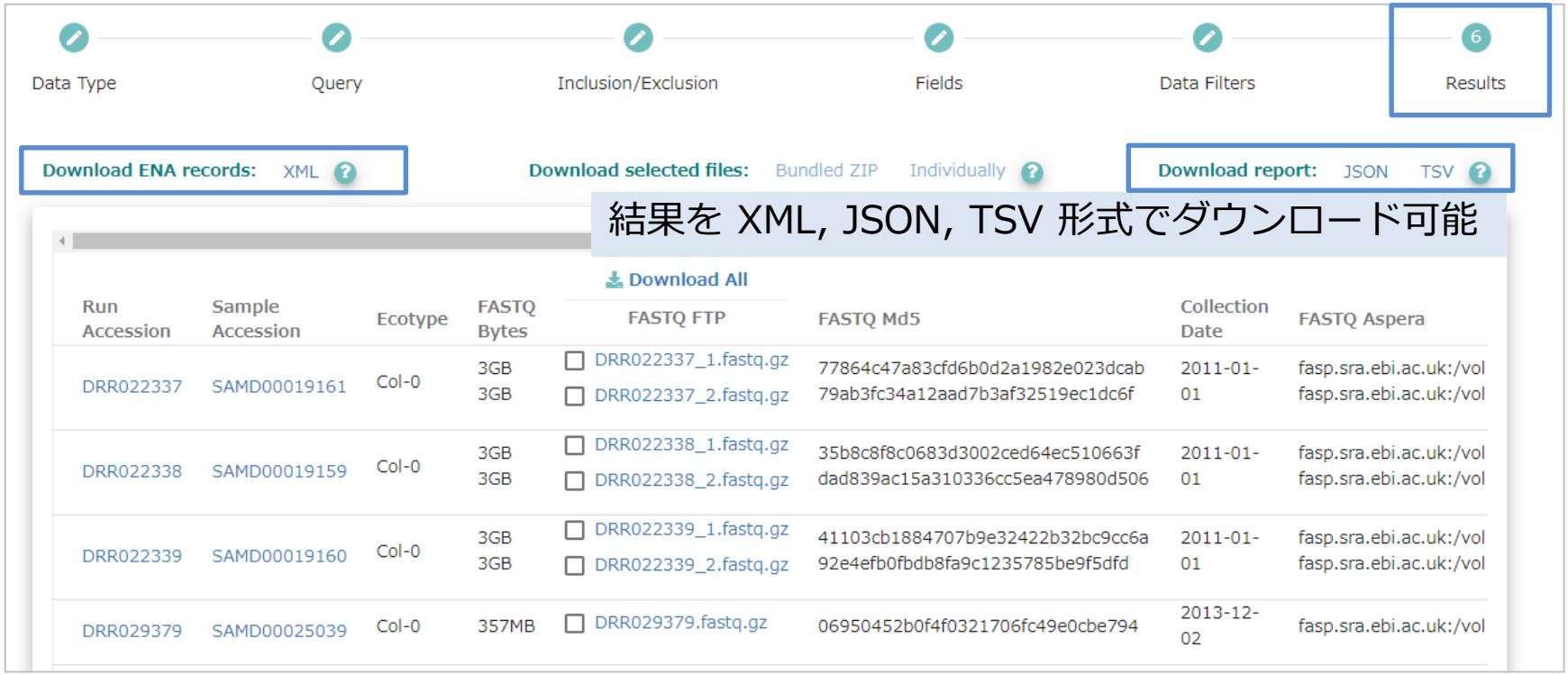
Advanced Search – Fields

出力項目を選択



Advanced Search – Results

結果を表形式で表示

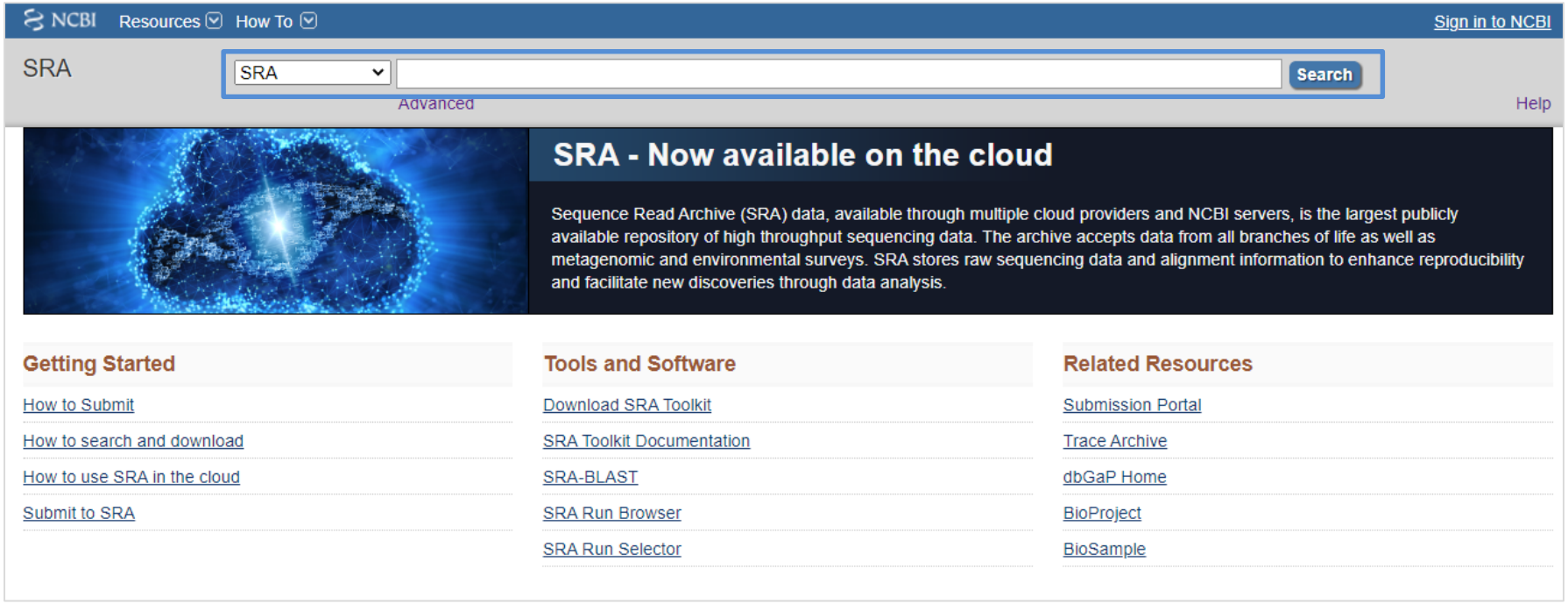


NCBI SRA Entrez

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

対象データベースを SRA に指定

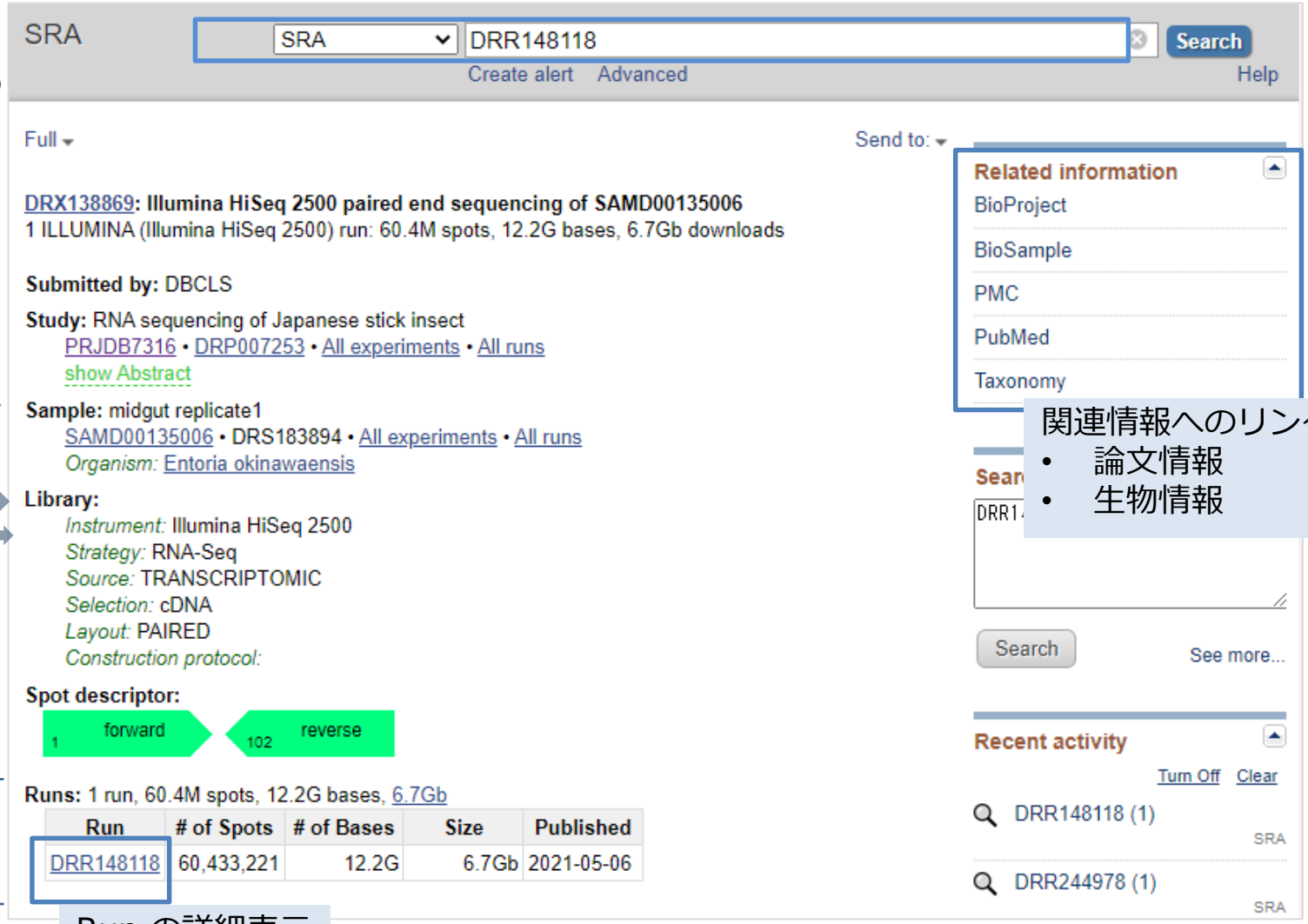
Query は Keyword, accession どちらも可



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

関連情報へのリンク
- 論文情報
- 生物情報

Run の詳細表示

https://www.ncbi.nlm.nih.gov/sra/?term=DRR148118

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

シークエンサー

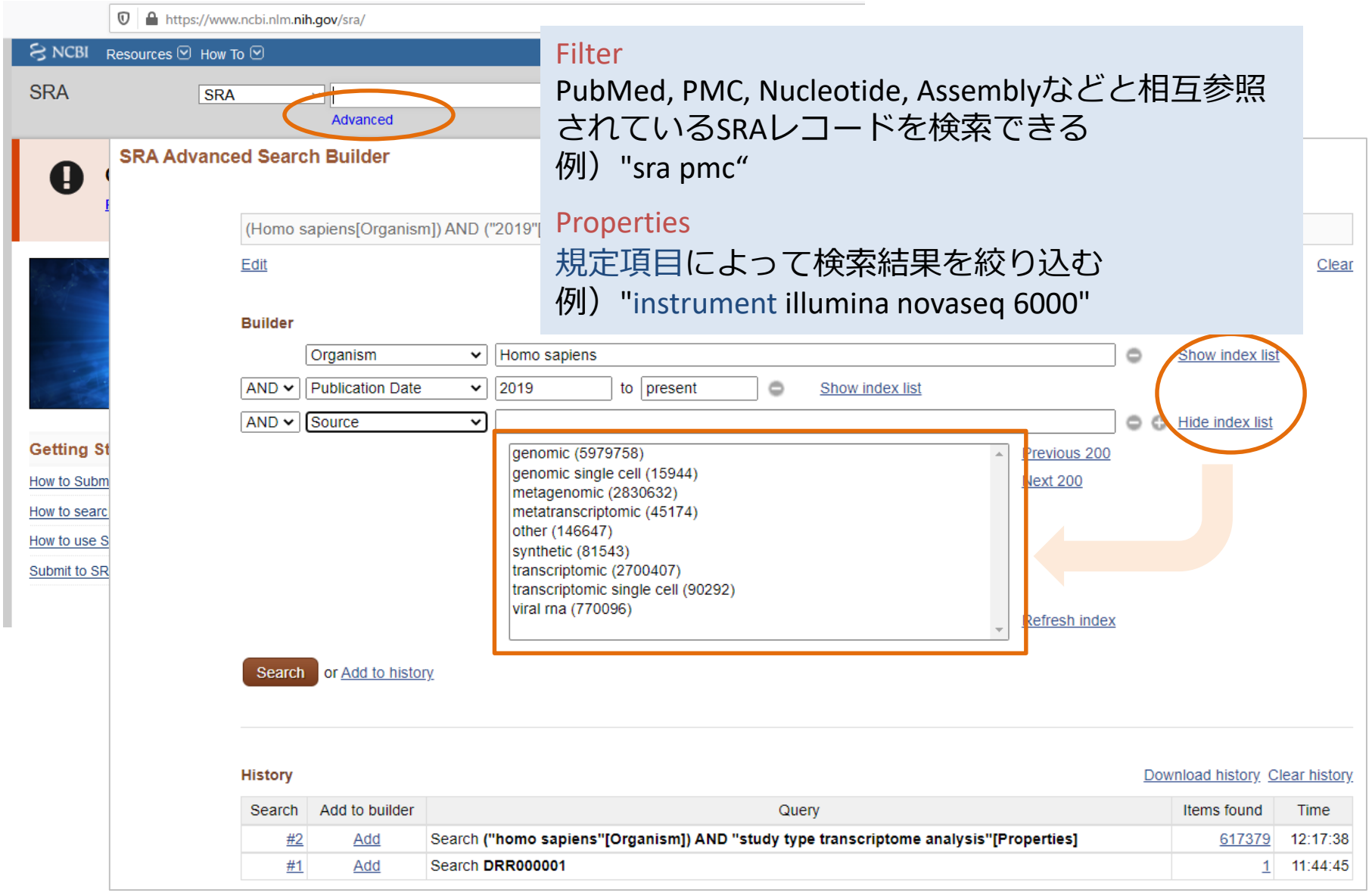
ライブラリーの構築手法

対象生物で絞り込み

関連DBで絞り込み

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

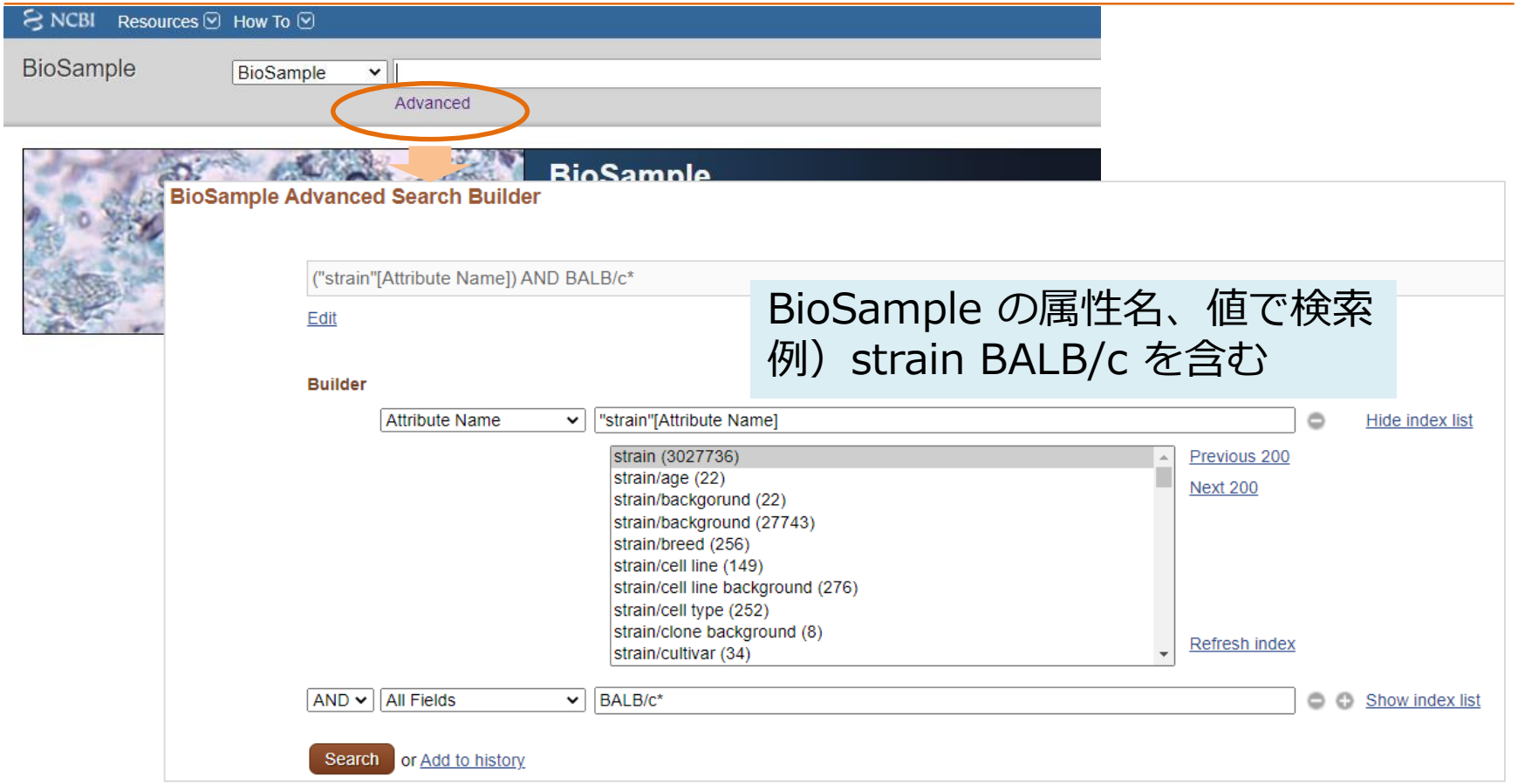
例）"sra pmc“

Properties

規定項目によって検索結果を絞り込む

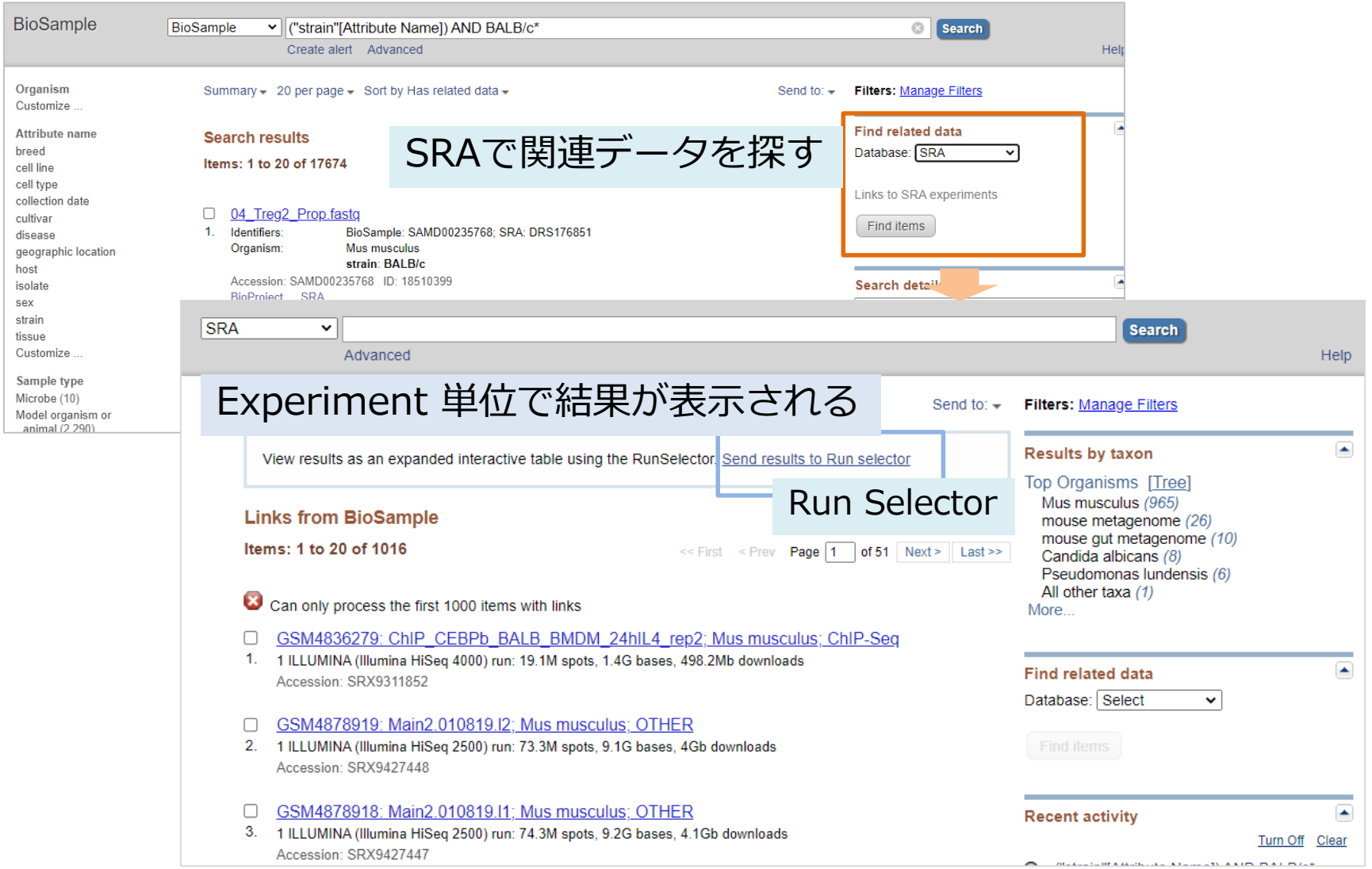
例）"instrument illumina novaseq 6000"

NCBI Advanced – BioSample



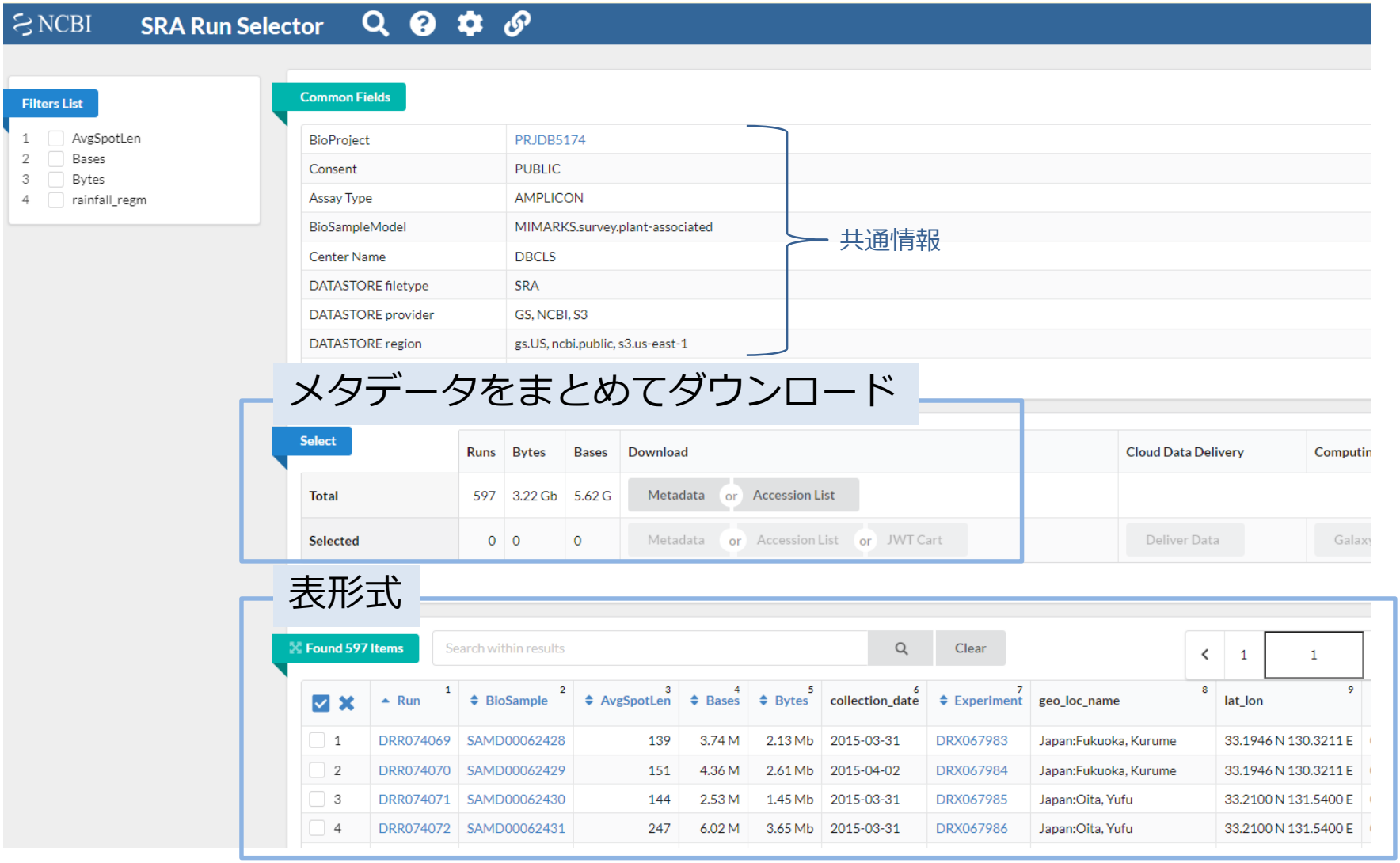
BioSample の属性名、値で検索
例）strain BALB/c を含む

NCBI Advanced –BioSample

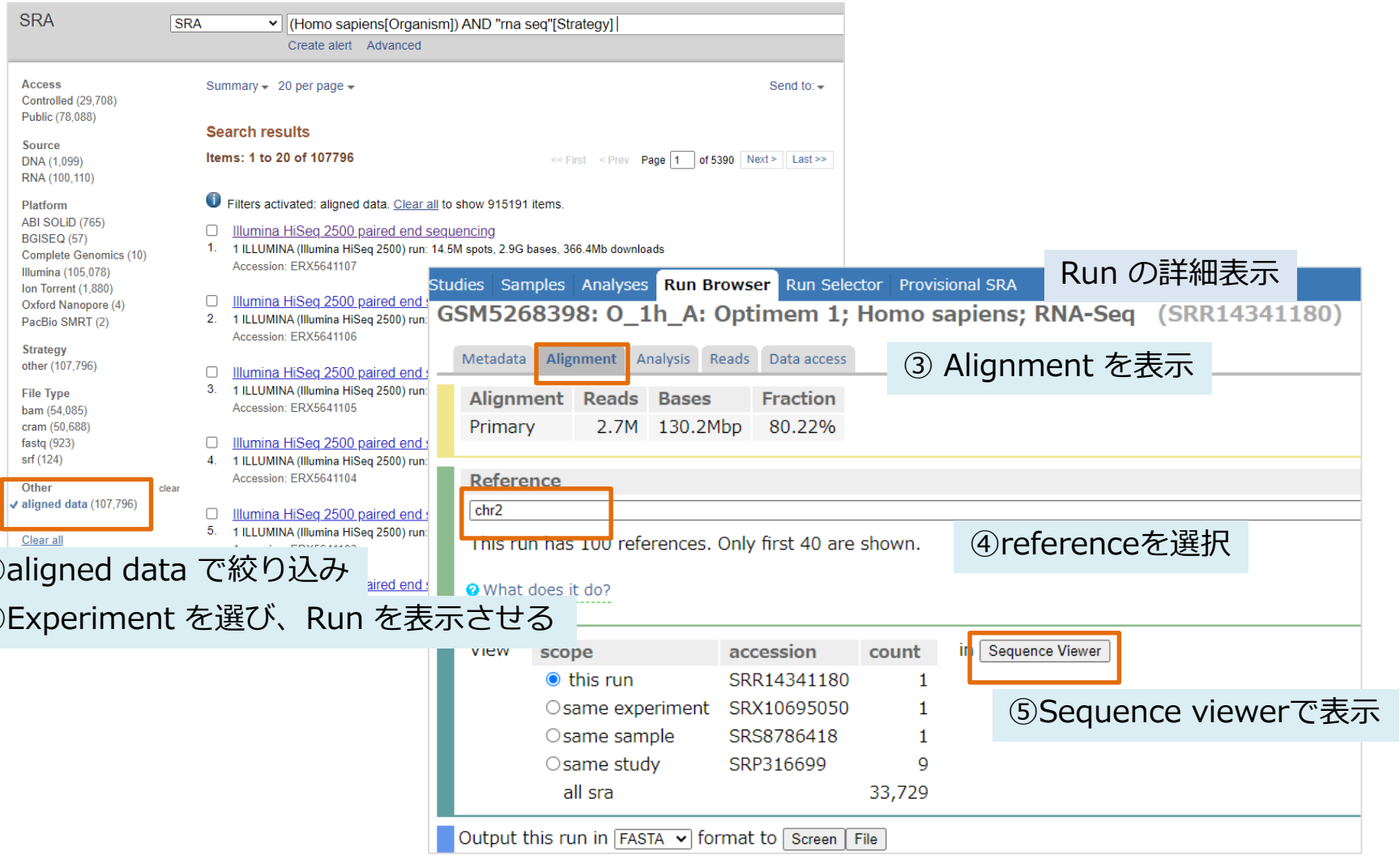


NCBI Run Selector

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



NCBI SRA aligned data の viewer 表示



NCBI - Sequence viewer

Run alignments

NCBI クラウドで SRAメタデータ検索

SRAデータは、パブリッククラウドにも移行されている

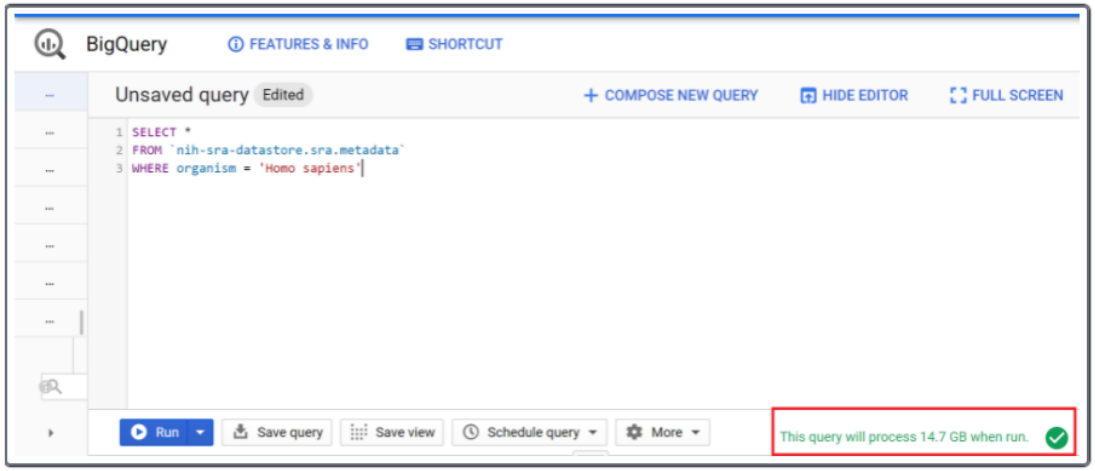
- Amazon Web Services（AWS）
- Google Cloud Platform（GCP）

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

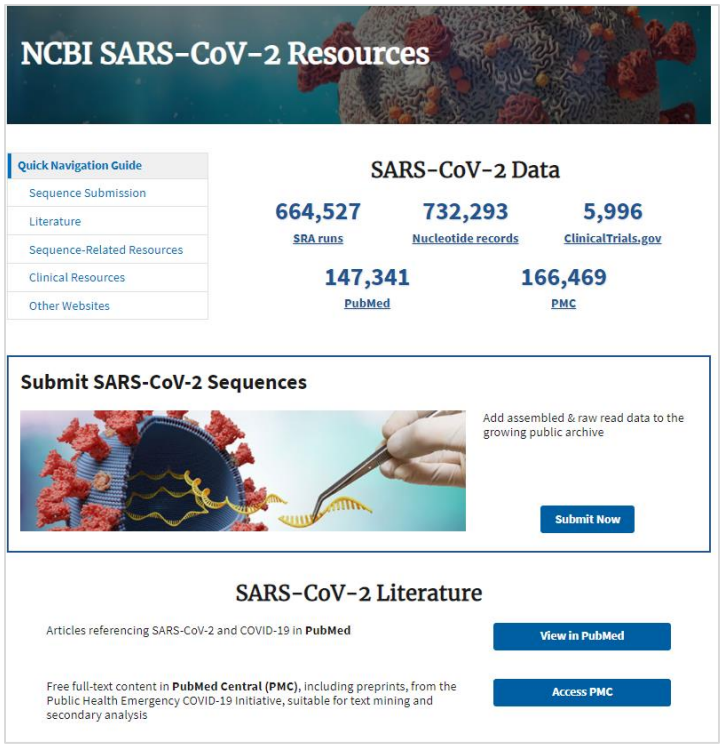


https://www.ncbi.nlm.nih.gov/sra/docs/sra-bigquery-examples/

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/



https://www.covid19dataportal.org/

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

データファイルの準備

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

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

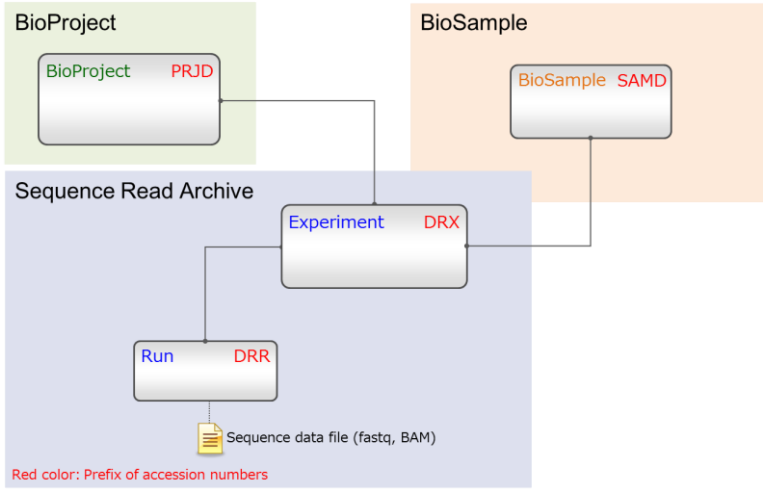
Run データファイル:
https://www.ddbj.nig.ac.jp/dra/submission.html#run-data-files

DRA登録に必要な主なメタデータ情報

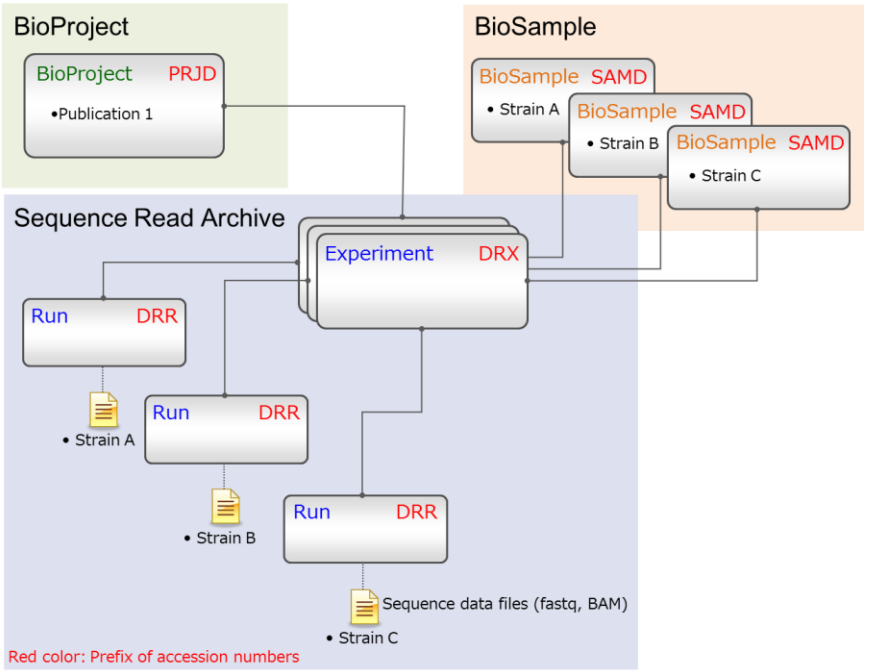
- 登録者情報
  - 登録内容に関して DDBJ からの問い合わせに対応できる方
- 公開予定
  - 即日公開 or 一定期間非公開
    - DRA: 公開予定日を 4 年以内で指定可能
- プロジェクト情報 (BioProject)
  - プロジェクトのタイトル、概要
  - 研究費情報
- 生物学的情報 (BioSample)
  - 生物名（NCBI Taxonomy database に登録されている種レベル以下の学名）
  - サンプルの採取環境、採取日時、特徴、処理条件など
- シークエンス用ライブラリーとシークエンスの手法 (Experiment)
  - ライブラリの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

BioSample の粒度

いくつのサンプルが必要か？

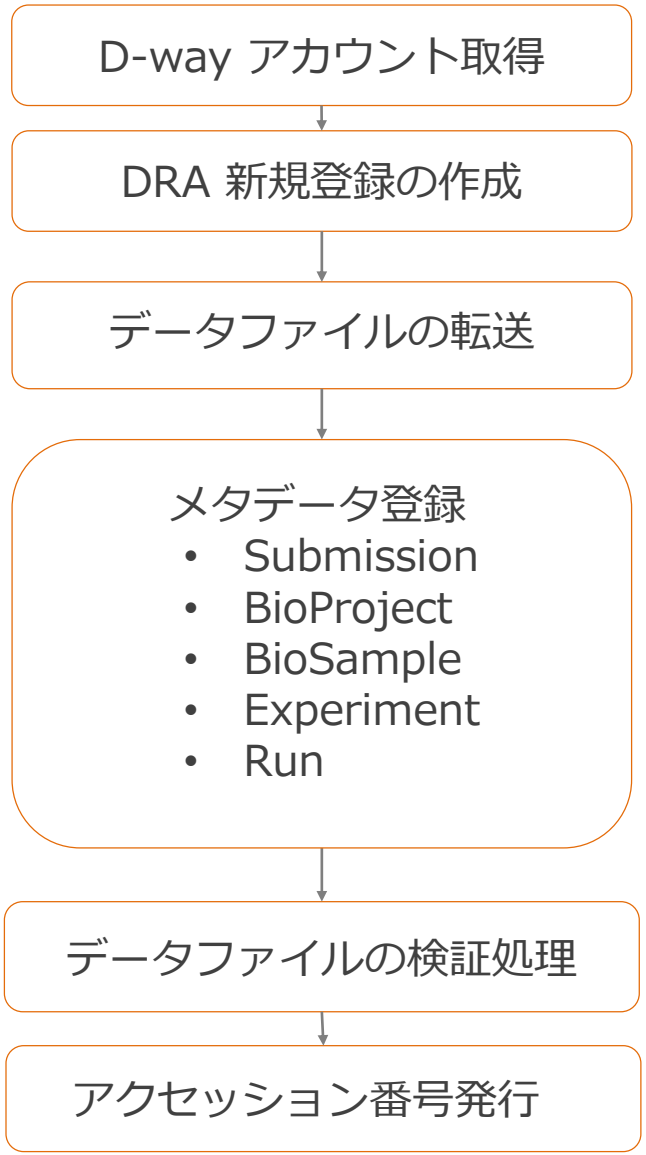
個別の BioSample として扱う場合の例

- ✓ 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

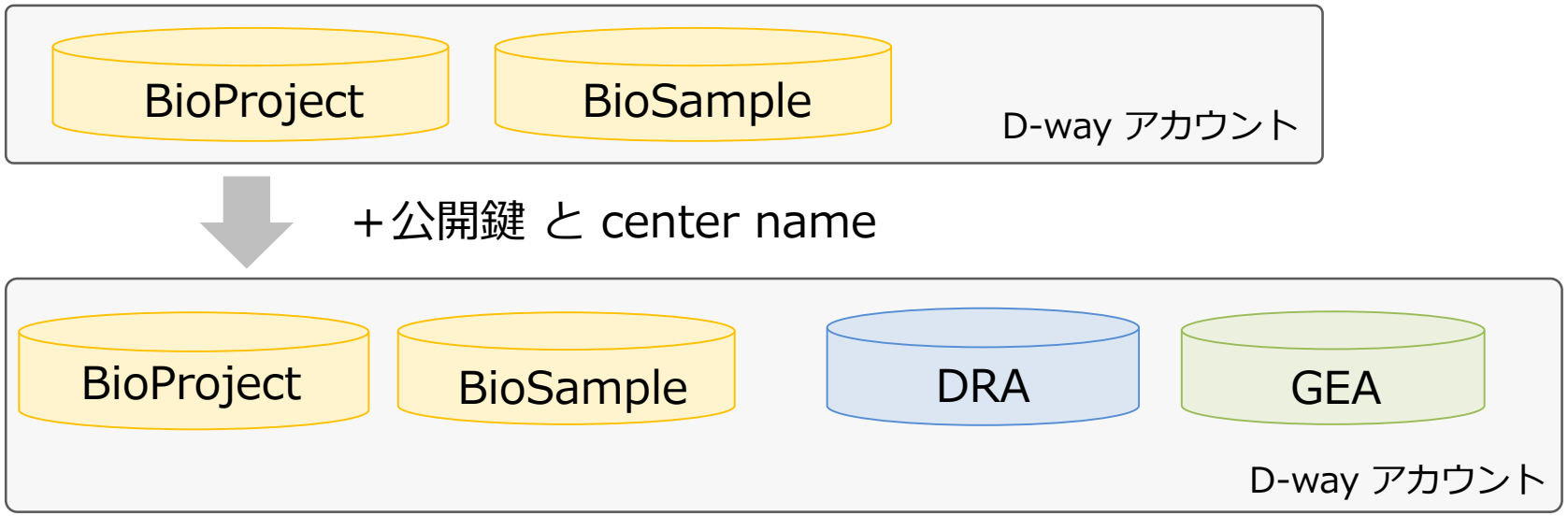


a) ウェブツールで作成
b) XML ファイルをアップロード

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

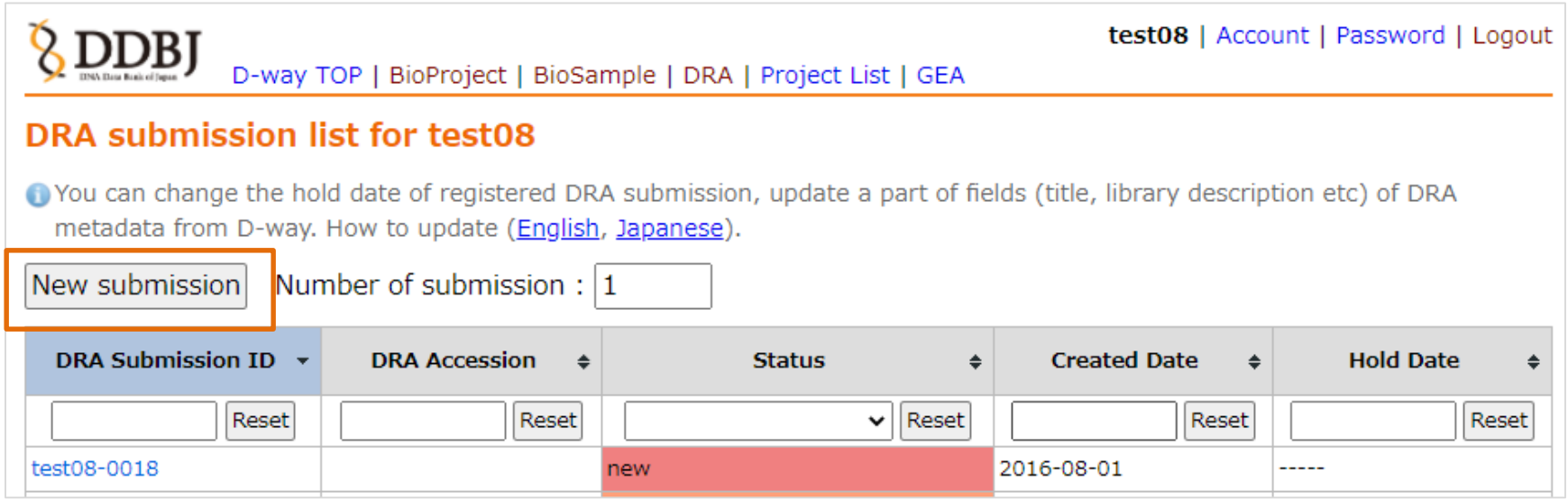
- ✓ 公開鍵と center name をアカウントに登録し、DRA 登録権限を取得



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

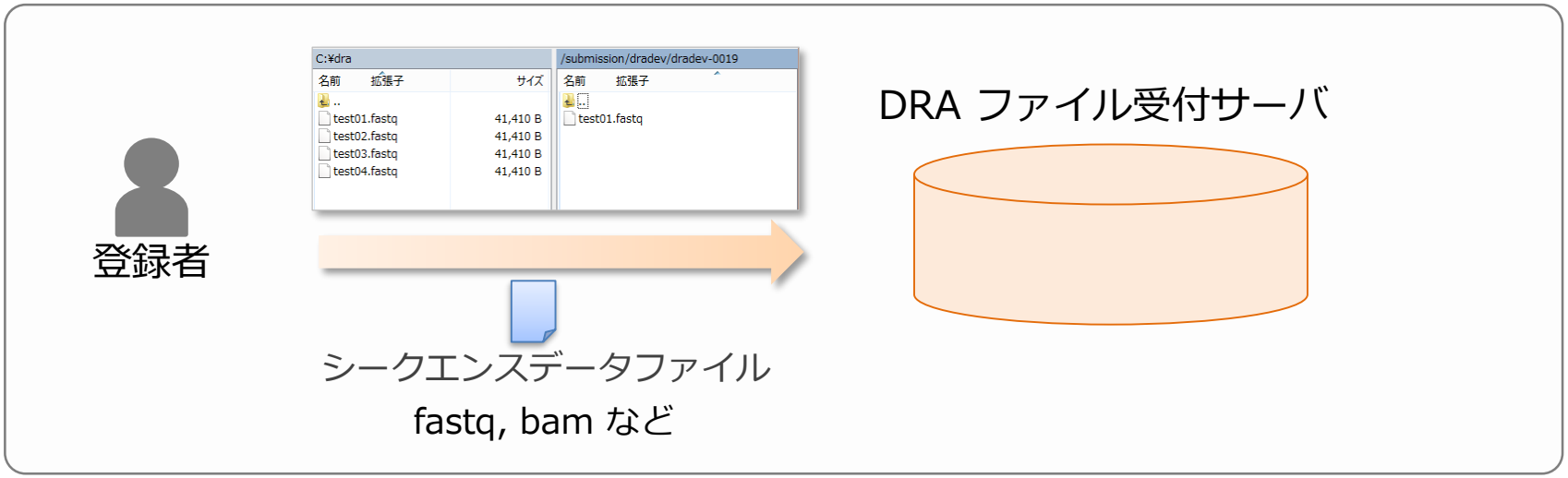
DRA Submission の作成



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



- ✓ 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

メタデータ登録

DDBJ DNA Data Bank of Japan

DRA Submission ID : test08-0017

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

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.a
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 int
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 uplo

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.

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)

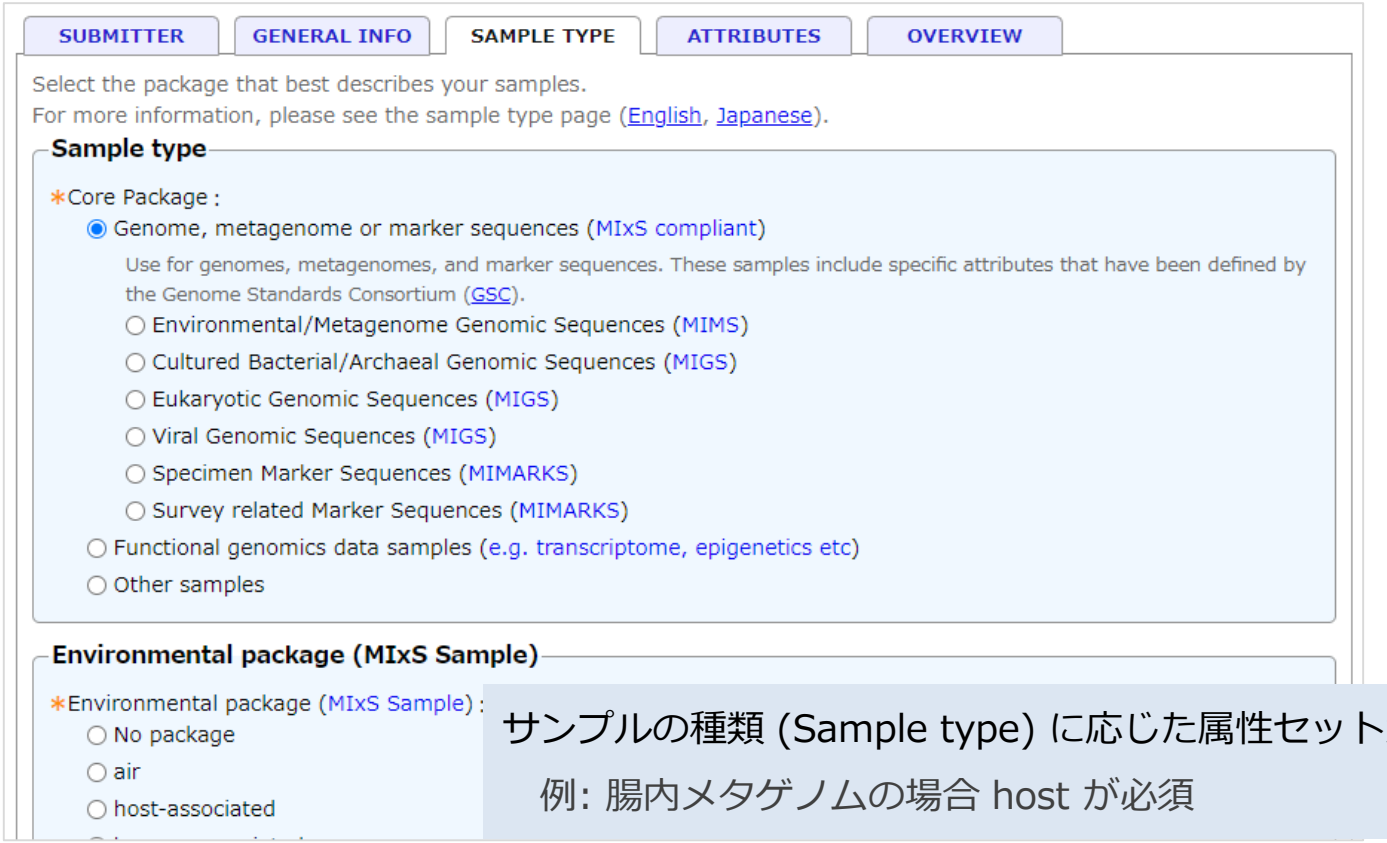
- ✓ 文献情報（Pubmed ID, DOI）

  # アクセッション番号を記載した論文情報を追加更新

BioSample 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）
- ✓ 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**



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)

*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

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

	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)

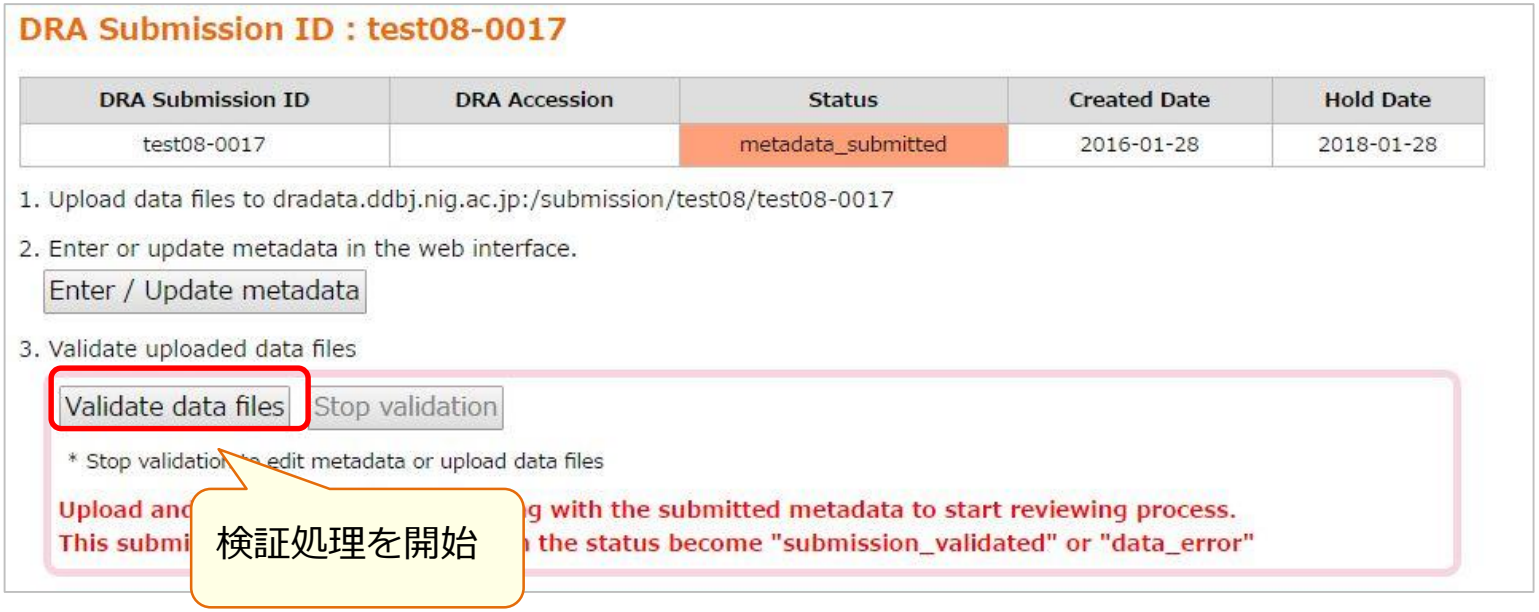
Run

✓ 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



✓ 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

アクセッション番号の発行

DDBJ DNA Data Bank of Japan

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

DRASearch　Send Feedback　Search Home　DRA Home

DRA000001 FTP

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

Navigation		
Study	DRP000001	
Experiment	DRX000001	FASTQ SRA
Sample	DRS000001	
Run	DRR000001	FASTQ SRA

ENA

Run: DRR000001

Illumina Genome Analyzer II paired end sequencing; B. subtilis subsp. natto genome sequencing September 2008

Organism:	Bacillus subtilis subsp. natto BEST195
Sample Accession:	SAMD00016353
Instrument Platform:	ILLUMINA
Instrument Model:	Illumina Genome Analyzer II
Read Count:	10148174
Base Count:	730668528
Center Name:	KEIO
Library Layout:	PAIRED
Library Strategy:	WGS
Library Source:	GENOMIC

Show More

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 FASTQ FTP
PRJDA38027	SAMD00016353	DRX000001	DRR000001	645657	Bacillus subtilis subsp. natto BEST195	DRR000001_1.fastq.gz DRR000001_2.fastq.gz

NCBI SRA

B. subtilis subsp. natto genome sequencing September 2008 (DRR000001)

Metadata　Reads　Download

Filter:　Find　Filtered Download　What does it do?

What can the filter be applied to?

< 1 1 1014818 >　View: biologic

1. DRR000001.1 DRS000001
name: 3060N:7:1:1116:340, member: default
x: 1116, y: 340

2. DRR000001.2 DRS000001
name: 3060N:7:1:1114:186, member: default
x: 1114, y: 186

3. DRR000001.3 DRS000001
name: 3060N:7:1:945:361, member: default
x: 945, y: 361

4. DRR000001.4 DRS000001
name: 3060N:7:1:1051:267, member: default
x: 1051, y: 267

5. DRR000001.5 DRS000001
name: 3060N:7:1:1074:240, member: default
x: 1074, y: 240

6. DRR000001.6 DRS000001
name: 3060N:7:1:914:322, member: default
x: 914, y: 322

7. DRR000001.7 DRS000001
name: 3060N:7:1:959:380, member: default
x: 959, y: 380

Reads (separated)

>gnl|SRA|DRR000001.1.1 3060N:7:1:1116:340 *(Biological)*
GATGGTAAGATAGAAGCAGTTGAAGTTTACAAACCG

>gnl|SRA|DRR000001.1.2 3060N:7:1:1116:340 *(Biological)*
GACACGTCCTTTTCATCATTCCACTTTGACTTGAAT