

# ゲノム時代の 生命情報・DDBJ センターが 果たしてきた役割

国立遺伝学研究所

生命情報・DDBJセンター(DDBJセンター)  
キュレータ 青野 英雄

# 講演内容

第1章 DDBJセンターの概要

第2章 塩基配列データベースの紹介  
(アノテーションを付与した/  
アセンブルした配列)

第3章 拡充されたデータベースの紹介

第4章 国立遺伝学研究所スパコンシステムの  
紹介

# 第1章

## DDBJセンターの概要

# DDBJセンターの紹介

**DNA Data Bank of Japan(DDBJ)** は、  
国立遺伝学研究所(静岡県三島市)に設置され、  
1987年より運営されています。



# INSDCの協力関係

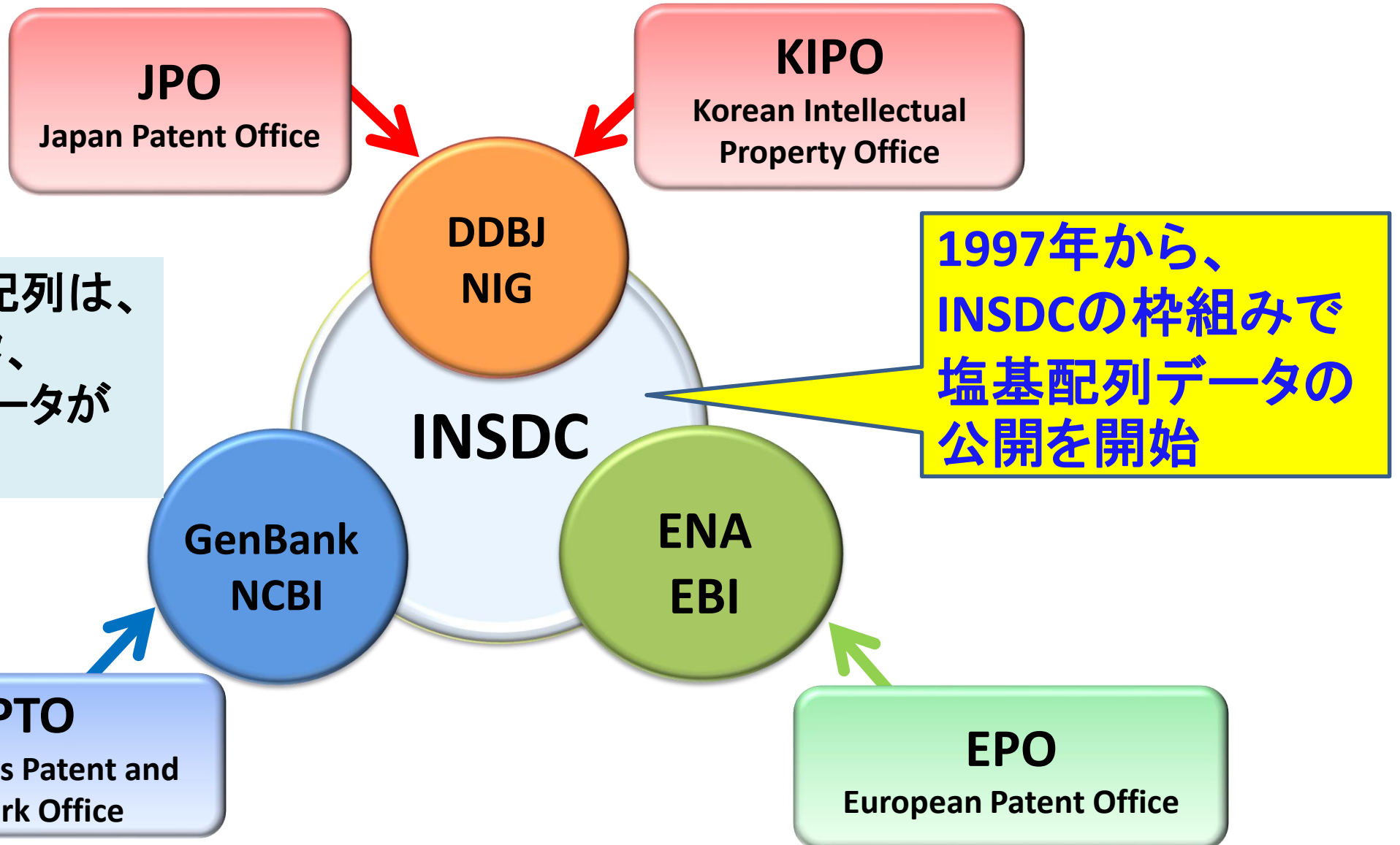
DDBJセンターは、欧州の **ENA/EBI**、および、米国の **NCBI** と共に、**国際塩基配列データベースコラボレーション**  
(INSDC: International Nucleotide Sequence Database Collaboration)  
を構築しています。



# INSDCで運用されるデータベース

	Annotated sequences	Capillary reads	NGS reads	Study	Sample	Assembly	Functional genomics	Variation	Genotype and phenotype
NCBI	GenBank	Trace Archive	Sequence Read Archive	BioProject	BioSample	Assembly	GEO	dbSNP/dbVar	dbGaP
EBI		European Nucleotide Archive (ENA)					ArrayExpress	EVA/DGVa	EGA
DDBJ	DDBJ	Trace Archive	Sequence Read Archive	BioProject	BioSample	Assembly	GEA	JVar-SNP/SV	JGA

# INSDCと特許庁の連携



特許出願関連配列は、塩基配列データ、アミノ酸配列データが含まれます。

# 第2章

塩基配列データベースの紹介  
(アノテーションを付与した/  
アセンブルした配列)

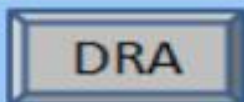
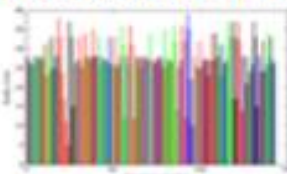


# Whole genome shotgun sequencing

Image data



Instrumentation data



Raw outputs  
(native or fastq files)

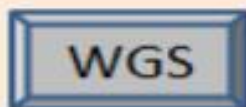
base calling

```
@Seq1
GGGTGATGGCCGCT ...
+Seq1
IIIIIIII9IG9IC ...
```

overlap detection



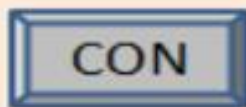
DDBJ



Contigs



assemble contigs



Scaffolds



finishing  
gap closure

general data  
(taxonomic)

Finished genomic sequences



アクセス制限データベース

JGA

個人レベルの遺伝型と表現型

NBDC

ヒトデータ審査委員会

プロジェクト情報

BioProject

BioSample

サンプル情報

説明  
箇所



# 配列データ登録の必要性

○配列データに基づく知見を論文に発表するためには、配列データを登録して発行されるアクセッション番号が必要となります。

○アクセッション番号は、その配列固有の番号となります。

Annotated/Assembled Data	
conventional	アルファベット 1 文字 + 5 桁の数字: 例 A12345 アルファベット 2 文字 + 6 桁の数字: 例 AB123456 アルファベット 2 文字 + 8 桁の数字: 例 AB12345678
bulk	アルファベット 4 文字 (For Large Scale Data) + 8 ~ 10 桁の数字: 例 ABCD01012345
WGS, TSA, TLS	アルファベット 6 文字 (For Large Scale Data) + 8 ~ 10 桁の数字: 例 ABCDEF01012345

# アノテーションを付与した/アセンブルした 塩基配列データの登録システム

DDBJ submission portal

## DDBJ Nucleotide Sequence Submission System (NSSS)

ウェブフォームの塩基配列登録システム

Nucleotide

Submission of small-scale nucleotide sequence data with annotation. In case of project data, please use BioProject, MSS, and DRA.

Create new submission

NSSS が対応していない多件数、長大な配列データの登録システム

Mass Submission System (MSS)

Please use mass submission system for the submission of following data. WGS, WGS scaffold(s), complete bacterial/eukaryotic genome, HTG, CON, GSS, EST, TSA, and other data includes huge number of sequences.



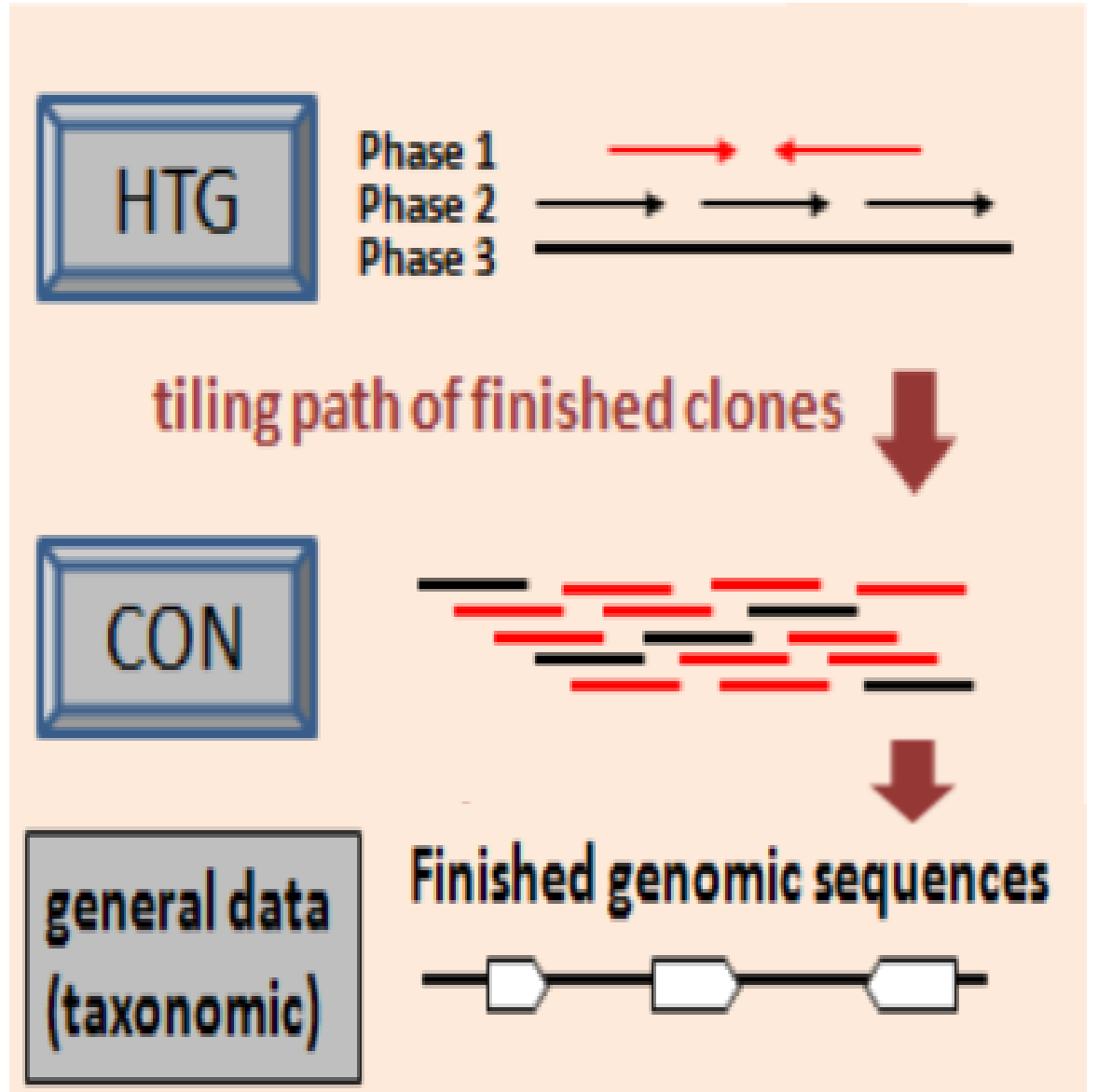
# ゲノムプロジェクト例

## ヒトゲノムプロジェクト

1990年に開始され、2003年に完成版が公開されました。

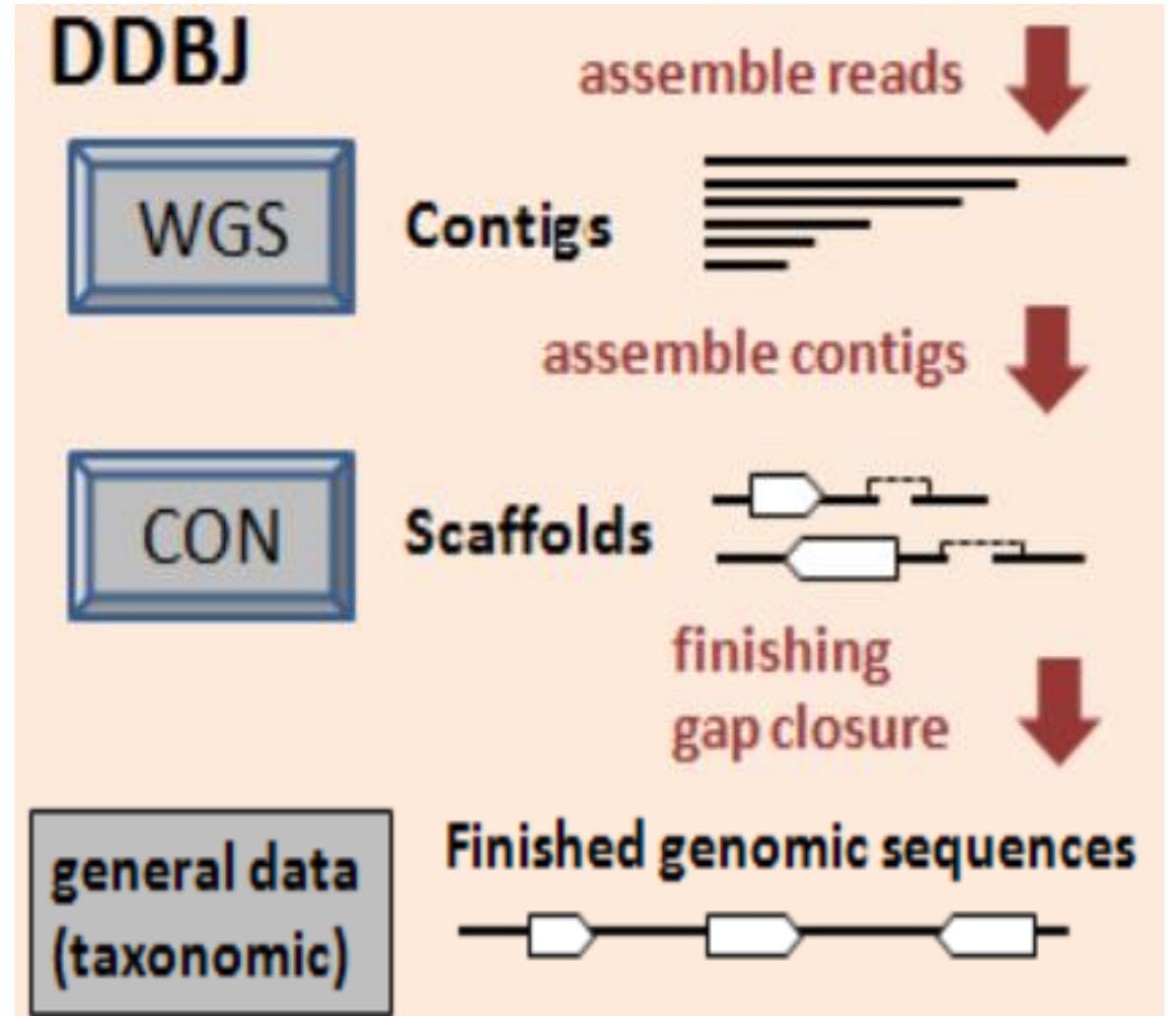
## ライスゲノムプロジェクト

2004年に国際イネゲノム解読プロジェクトによって、  
*Oryza sativa* ssp. *japonica* cultivar Nipponbare  
ゲノムが完全解読されました。



# WGS ( Whole Genome Shotgun )

WGS法は、ゲノム全体を物理的に断片化し、シーケンサで各断片の塩基配列を決定した後、プログラムを用いてアセンブルして完成させます。





# WGSデータ例


## マレーシアの ヒトから採取された Helicobacter pylori UM045 の WGSデータ

```
LOCUS      AON001000001                20372 bp    DNA        linear    BCT 24-MAY-2013
DEFINITION Helicobacter pylori UM045 1, whole genome shotgun sequence.
ACCESSION  AON001000001 AON001000000
VERSION    AON001000001.1
DBLINK     BioProject: PRJNA187438
           BioSample: SAMN02471801
KEYWORDS   WGS.
SOURCE     Helicobacter pylori UM045
   ORGANISM Helicobacter pylori UM045
           Bacteria; Proteobacteria; Epsilonproteobacteria; Campylobacterales;
           Helicobacteraceae; Helicobacter.
REFERENCE  1 (bases 1 to 20372)
   AUTHORS Kumar,N., Baddam,R., Mariappan,V., Shaik,S., Tiruvayipati,S.,
           Tenguria,S., Goh,K.L., Vadivelu,J. and Ahmed,N.
   TITLE    Comparative Genomics of Helicobacter pylori isolates obtained from
           different ethnic groups in Malaysia
   JOURNAL   Unpublished
REFERENCE  2 (bases 1 to 20372)
   AUTHORS Kumar,N., Baddam,R., Mariappan,V., Shaik,S., Tiruvayipati,S.,
           Tenguria,S., Goh,K.L., Vadivelu,J. and Ahmed,N.
   TITLE    Direct Submission
   JOURNAL   Submitted (30-JAN-2013) School of Life sciences, Department of
           Biotechnology, University of Hyderabad, Prof. C. R. Rao Road,
           Gachibowli, Hyderabad, Andhra Pradesh 500046, India
COMMENT    Bacteria and source DNA available from: Prof. Dr. Jamunarani A/P S
           Vadivelu, Department of Medical Microbiology, Faculty of Medicine
           Building, University of Malaya, 50603 Kuala Lumpur, Malaysia.

           ##Genome-Assembly-Data-START##
           Assembly Method      :: Velvet v. 1.2.03
           Genome Coverage      :: 200.0x
           Sequencing Technology :: Illumina GAIIx
           ##Genome-Assembly-Data-END##
FEATURES   Location/Qualifiers
   source   1..20372
            /organism="Helicobacter pylori UM045"
            /mol_type="genomic DNA"
            /strain="UM045"
            /host="Homo sapiens"
            /db_xref="taxon:1287065"
            /country="Malaysia"
BASE COUNT 6522 a          4435 c          3810 g          5605 t
ORIGIN      1 tttttttttt ttgtcatatt tcttaaaaat tttagaat ttgtgatttt ggtgtttttt
           省略
//
```

# DDBJ Fast Annotation and Submission Tool (DFAST)

DFASTは**原核生物ゲノムの自動アノテーションサービス**です。  
出力結果をDDBJのMSS登録に利用することができます。



## DDBJ Fast Annotation and Submission Tool

[Start your project!](#) [ Running 0 / Waiting 0 ]

Please see [FAQ](#) and [Sample Result](#) if this is your first visit.

DFAST Legacy server (based on Prokka).

Organism-specific reference databases (manually curated)

[Lactic Acid Bacteria](#) (1.0) [Bifidobacterium](#) (β0.1) [Cyanobacteria](#) (β0.1) [E. coli](#) (β0.1)

General-purpose reference databases

[RefSeq](#) (1.0, automatically curated using protein sequences mainly from 'Reference Genomes' in RefSeq.)

Subsets for following phyla are also available: [Actinobacteria](#) [Firmicutes](#) [Proteobacteria](#)

### DAGA : DFAST Archive of Genome Annotation

DAGA stores genomic data collected from the public nucleotide database and the sequence read archive. All the genomes are consistently annotated using DFAST.  
Currently, DAGA is available only for genomes of Lactic Acid Bacteria.

1421 annotated genome resources are available, covering 2 genera and 191 species. [ENTER](#)

### DFAST Prokaryotic genome annotation pipeline

Query File (Fasta format, up to 15Mbyte)

[参照...](#) ファイルが選択されていません。 ☐ Run in demo mode (Sample annotation for E.coli O26)

Job Title

(optional)

Mail Address

E-mail notification will be sent to this address when the job is completed. (optional)

--- options ---

Additional DB

Minimum sequence length

Genetic code ☒ 11 (Standard bacterial/archaeal) ☐ 4 (Mycoplasma/Spiroplasma)

CDS prediction tool ☒ MetaGeneAnnotator (default) ☐ Prodigal

tRNA prediction tool ☒ Aragorn (default) ☐ tRNAscan-SE (Bacteria) ☐ tRNAscan-SE (Archaea)

☐ Enable HMM scan against TIGRFAM ☐ Enable RPSBLAST against COG

☐ Sort sequences by length (the longer comes first)

Fix the sequence origin (only for a finished genome with a circular chromosome)

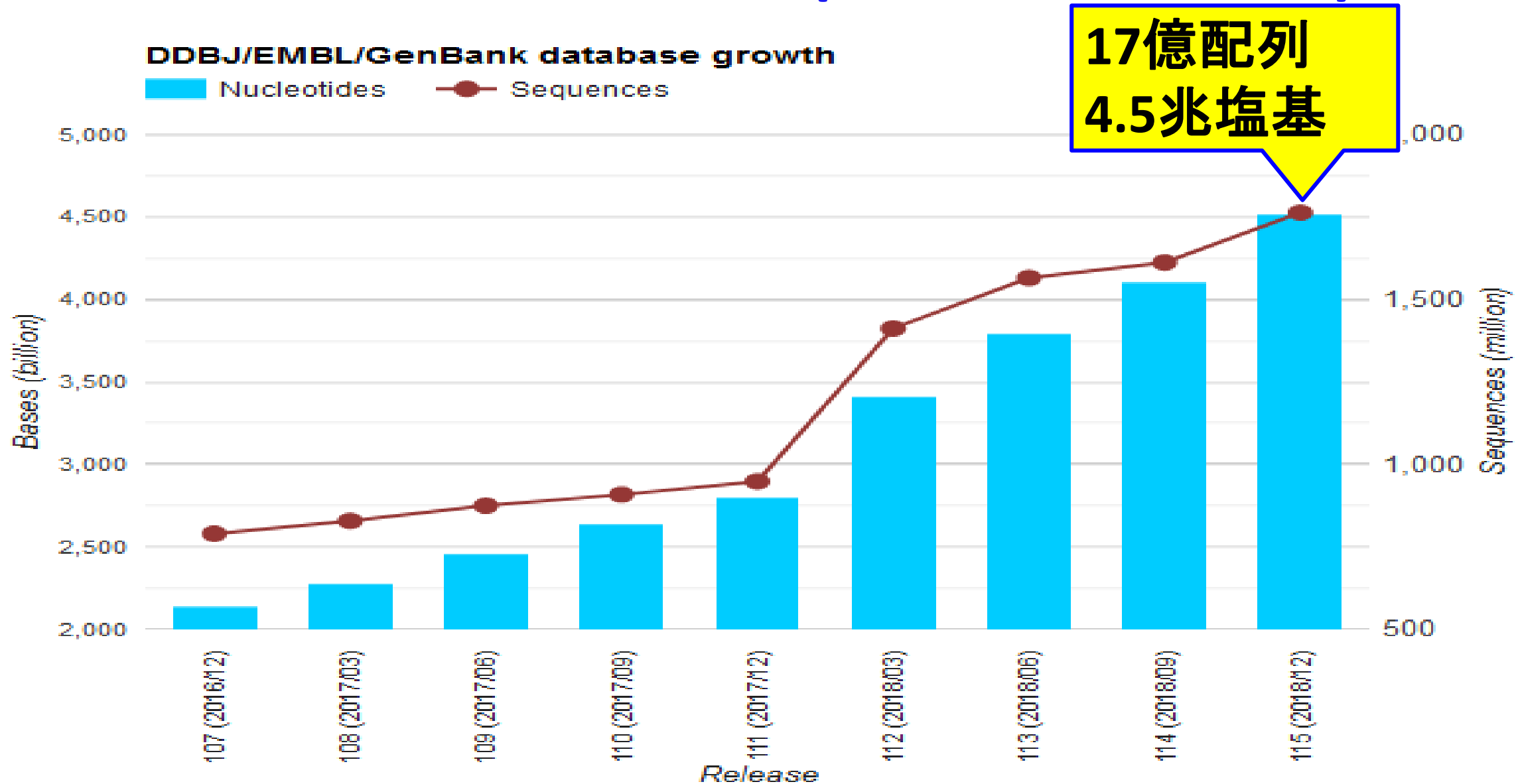
☐ Rotate/flip the chromosome so that the dnaA gene comes first

Offset from the start codon of the dnaA gene:  bp

[Run](#)



# DDBJ 総データ量 (塩基数/配列数)



# データ検索サービス

getentry

アクセッション番号による  
検索サービス

アクセッション番号等によるエントリ検索

ID :

検索

DNA データベース : ☒ DDBJ / EMBL / GenBank ☐ MGA

出力形式 : フラットファイル(DDBJ) ▾

Protein データベース : ☐ UniProt ☐ PDB ☐ DAD ☐ Patent

出力形式 : default ▾

取得方法 :

html ▾

上限 : 10 件

ARSA (Search Condition)

キーワードによる  
検索サービス

Quick Search

Search

AND ▾

# 第3章

## 拡充された

## データベースの紹介

次世代シーケンサーの登場により、

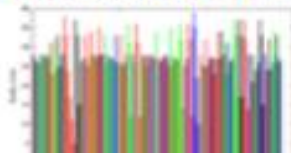
データの受け入れのために、新規データベースの開発と拡充を行ってきました。

## Whole genome shotgun sequencing

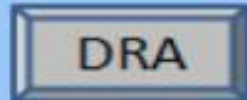
Image data



Instrumentation data



base calling



Raw outputs  
(native or fastq files)

```
@Seq1
GGGTGATGGCCGCT ...
+Seq1
IIIIIIII9IG9IC ...
```

overlap detection

説明  
箇所



アクセス制限データベース

JGA

個人レベルの遺伝型と表現型

NBDC

ヒトデータ審査委員会

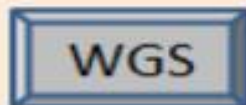
プロジェクト情報

**BioProject**

**BioSample**

サンプル情報

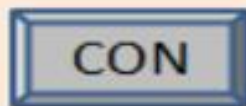
DDBJ



Contigs



assemble contigs



Scaffolds



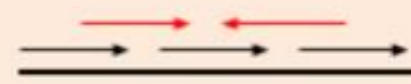
finishing  
gap closure

general data  
(taxonomic)

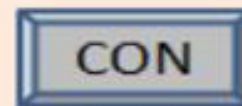
Finished genomic sequences



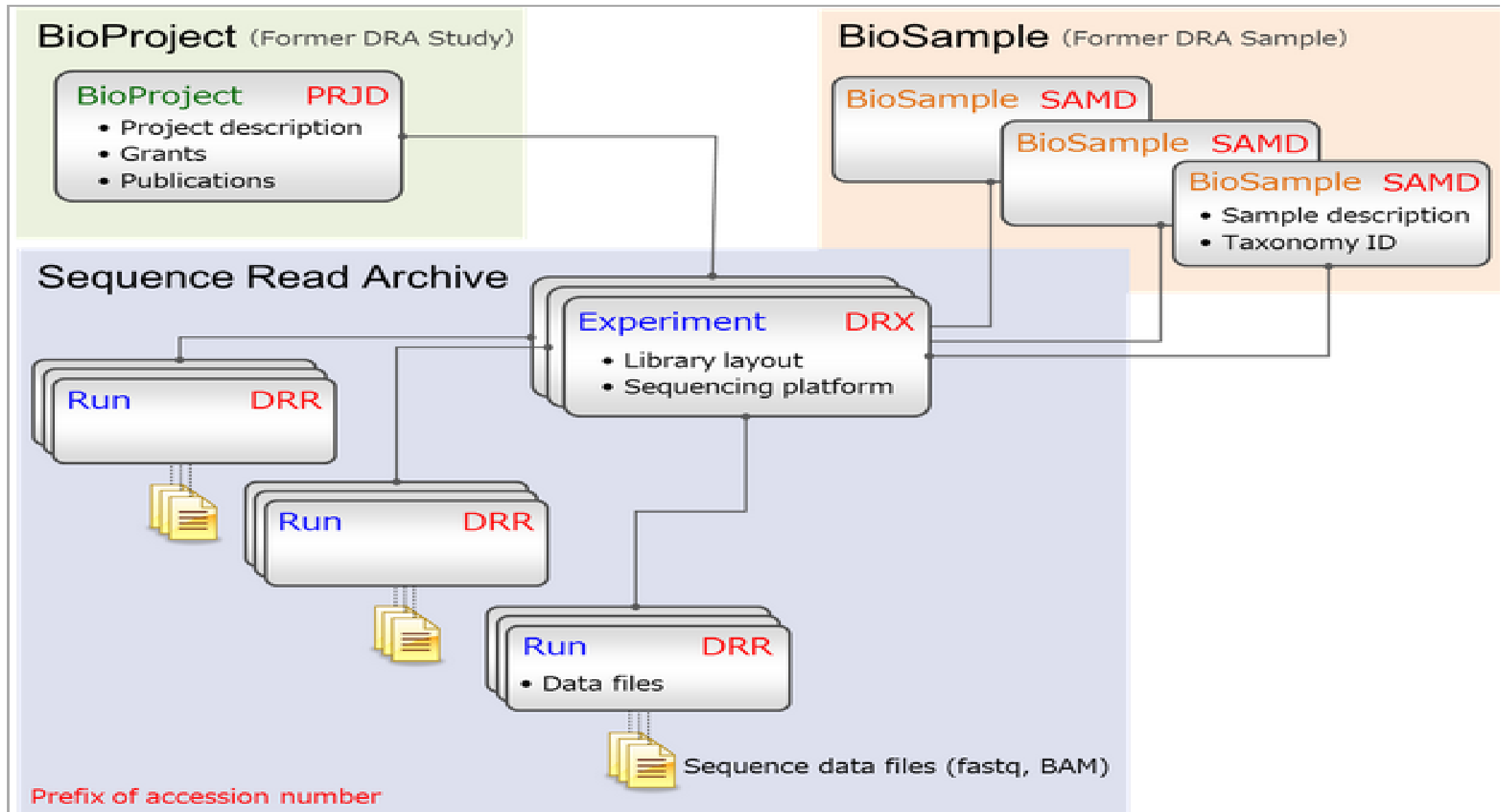
Phase 1  
Phase 2  
Phase 3



tiling path of finished clones



# DRA-BioProject-BioSample 連携図



# DDBJ Sequence Read Archive(DRA)

次世代シーケンサからの出力データとアライメントデータの  
SRA (Sequence Read Archive)データベース。

データ例: DRA000032

Submission Detail		Navigation	
Alias	DRA000032	Study	DRP000032
Submission ID		Experiment	DRX000056 <a href="#">FASTQ</a> <a href="#">SRA</a>
Submission Date	2009-05-08	Sample	DRS000055
Center Name	UT-MGS	Run	DRR000119 <a href="#">FASTQ</a> <a href="#">SRA</a>
Lab Name	Laboratory of Functional Genomics, Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo		DRR000120 <a href="#">FASTQ</a> <a href="#">SRA</a>
			DRR000121 <a href="#">FASTQ</a> <a href="#">SRA</a>
			DRR000122 <a href="#">FASTQ</a> <a href="#">SRA</a>

Experiment Detail	
Title	mouse_embryo_7d_TSS
Design Description	Transcriptional start site analysis
Organism	Mus musculus
Library Description	
Name	mouse_embryo_7day
Strategy	FL-cDNA
Source	TRANSCRIPTOMIC
Selection	cDNA
Layout	SINGLE
Construction Protocol	none provided

Run Detail	
Alias	DRR000119
Instrument model	
Date of run	2008-10-07
Run center	UT-MGS
Number of spots	7,374,187
Number of bases	265,470,732
READS (joined) <span>quality <input type="checkbox"/> show 10 rows</span>	
>DRR000119.1	
CCTTCTCCTTCGACCCCCGCGATCTCCACTCTTTCC	

# BioProject

研究プロジェクトの関連するデータをまとめたプロジェクトのデータベース。

プロジェクト  
概要



生物情報  
論文情報



関連配列  
情報



**Ptychodera flava** Accession: PRJDB3182 ID: 302624

**Whole Genome Shotgun Sequencing of an Hawaiian Acornworm, *Ptychodera flava*.**

Whole Genome Shotgun Sequencing of an Hawaiian Acornworm, *Ptychodera flava*. All original genomic library was produced from sperm spawned from one male individual which was sampled near Oahu island in Hawaii at 2006 Dec. *Ptychodera flava* is an acornworm which shows indirect development so that it is very useful model organisms to study the evolution of Deuterostomia. DRA002855 is the accession number for sequences. [Less...](#)

Accession	PRJDB3182
Data Type	Genome sequencing and assembly
Scope	Monoisolate
Organism	<b><i>Ptychodera flava</i></b> [Taxonomy ID: 63121] Eukaryota; Metazoa; Hemichordata; Enteropneusta; Ptychoderidae; Ptychodera; <i>Ptychodera flava</i>
Publications	1. Simakov O <i>et al.</i> , "Hemichordate genomes and deuterostome origins.", <i>Nature</i> , 2015 Nov 18;527(7579):459-65 2. Published online, DOI: 10.1038/nature16150
Submission	Registration date: 24-Nov-2015 Okinawa Institute of Science and Technology
Related Resources	<ul style="list-style-type: none"><li>• <a href="#">Marine Genomics Unit, OIST</a></li><li>• <a href="#">Marine Biological Laboratory (MBL), Graduate School of Science, Hiroshima University</a></li><li>• <a href="#">DRA002855</a></li><li>• <a href="#">BCFJ01000001-BCFJ01317432</a></li></ul>
Relevance	Evolution

[See Genome Information for \*Ptychodera flava\*](#)

[NAVIGATE ACROSS](#)  
3 additional projects are related by organism.

**Project Data:**

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	218256
WGS master	1
SRA Experiments	11
PUBLICATIONS	
PubMed	1
PMC	1
OTHER DATASETS	
BioSample	1
Assembly	1

**Assembly details:**

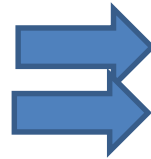
Assembly	Level	WGS	BioSample	Taxonomy
GCA_001465055.1	Scaffold	BCFJ000000000	SAMD00023482	<i>Ptychodera flava</i>

Download

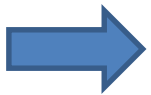
# BioSample

生物学的な試料(サンプル)の情報を集中管理するデータベース。

生物情報  
パッケージ



サンプル属性



関連データ



Ptychodera flava Genomic DNAs																																						
Identifiers	BioSample: SAMD00023482																																					
Organism	<a href="#">Ptychodera flava</a> cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Hemichordata; Enteropneusta; Ptychoderidae; Ptychodera																																					
Package	<a href="#">MIGS: eukaryote: version 4.0</a>																																					
Attributes	<table><tr><td><b>sample name</b></td><td>Ptychodera flava Genomic DNAs</td></tr><tr><td><b>collection date</b></td><td>2006-12-10</td></tr><tr><td><b>broad-scale environmental context</b></td><td>sea</td></tr><tr><td><b>local environmental context</b></td><td>sand</td></tr><tr><td><b>environmental medium</b></td><td>sea water</td></tr><tr><td><b>geographic location</b></td><td><a href="#">USA: HI, Oahu</a></td></tr><tr><td><b>latitude and longitude</b></td><td>NA</td></tr><tr><td><b>project name</b></td><td>Ptychodera flava Genomic DNAs</td></tr><tr><td><b>isolation and growth condition</b></td><td>NA</td></tr><tr><td><b>reference for biomaterial</b></td><td>NA</td></tr><tr><td><b>number of replicons</b></td><td>NA</td></tr><tr><td><b>estimated size</b></td><td>800,000,000</td></tr><tr><td><b>ploidy</b></td><td>haploid</td></tr><tr><td><b>propagation</b></td><td>NA</td></tr><tr><td><b>cultivar</b></td><td>missing</td></tr><tr><td><b>ecotype</b></td><td>missing</td></tr><tr><td><b>isolate</b></td><td>missing</td></tr><tr><td><b>strain</b></td><td>missing</td></tr></table>		<b>sample name</b>	Ptychodera flava Genomic DNAs	<b>collection date</b>	2006-12-10	<b>broad-scale environmental context</b>	sea	<b>local environmental context</b>	sand	<b>environmental medium</b>	sea water	<b>geographic location</b>	<a href="#">USA: HI, Oahu</a>	<b>latitude and longitude</b>	NA	<b>project name</b>	Ptychodera flava Genomic DNAs	<b>isolation and growth condition</b>	NA	<b>reference for biomaterial</b>	NA	<b>number of replicons</b>	NA	<b>estimated size</b>	800,000,000	<b>ploidy</b>	haploid	<b>propagation</b>	NA	<b>cultivar</b>	missing	<b>ecotype</b>	missing	<b>isolate</b>	missing	<b>strain</b>	missing
<b>sample name</b>	Ptychodera flava Genomic DNAs																																					
<b>collection date</b>	2006-12-10																																					
<b>broad-scale environmental context</b>	sea																																					
<b>local environmental context</b>	sand																																					
<b>environmental medium</b>	sea water																																					
<b>geographic location</b>	<a href="#">USA: HI, Oahu</a>																																					
<b>latitude and longitude</b>	NA																																					
<b>project name</b>	Ptychodera flava Genomic DNAs																																					
<b>isolation and growth condition</b>	NA																																					
<b>reference for biomaterial</b>	NA																																					
<b>number of replicons</b>	NA																																					
<b>estimated size</b>	800,000,000																																					
<b>ploidy</b>	haploid																																					
<b>propagation</b>	NA																																					
<b>cultivar</b>	missing																																					
<b>ecotype</b>	missing																																					
<b>isolate</b>	missing																																					
<b>strain</b>	missing																																					
Description	Hawaiian Acornworm, Ptychodera flava, Genomic DNA Keywords: GSC:MIxS;MIGS:4.0																																					
Links	<a href="#">marinegenomicsdb</a>																																					
BioProject	<a href="#">PRJDB3182</a> Ptychodera flava Retrieve <a href="#">all samples</a> from this project																																					
Submission	<a href="#">Okinawa Institute of Science and Technology Marine Genomics Unit</a> ; 2015-11-17																																					
Accession: SAMD00023482 ID: 4272872 <a href="#">BioProject</a> <a href="#">SRA</a> <a href="#">Nucleotide</a>																																						



# DRAデータの登録の流れ

D-way 登録アカウントを作成



データファイルをアップロード



## プロジェクトとサンプル情報を登録

**BioProject (Study):** 研究プロジェクトの情報を登録

**BioSample (Sample):** 生物学的、物理的にユニークなサンプル情報を登録



## Experiment と Run の登録と検証処理

**DRA Experiment:**

特定のサンプルから構築したライブラリーについての情報を登録

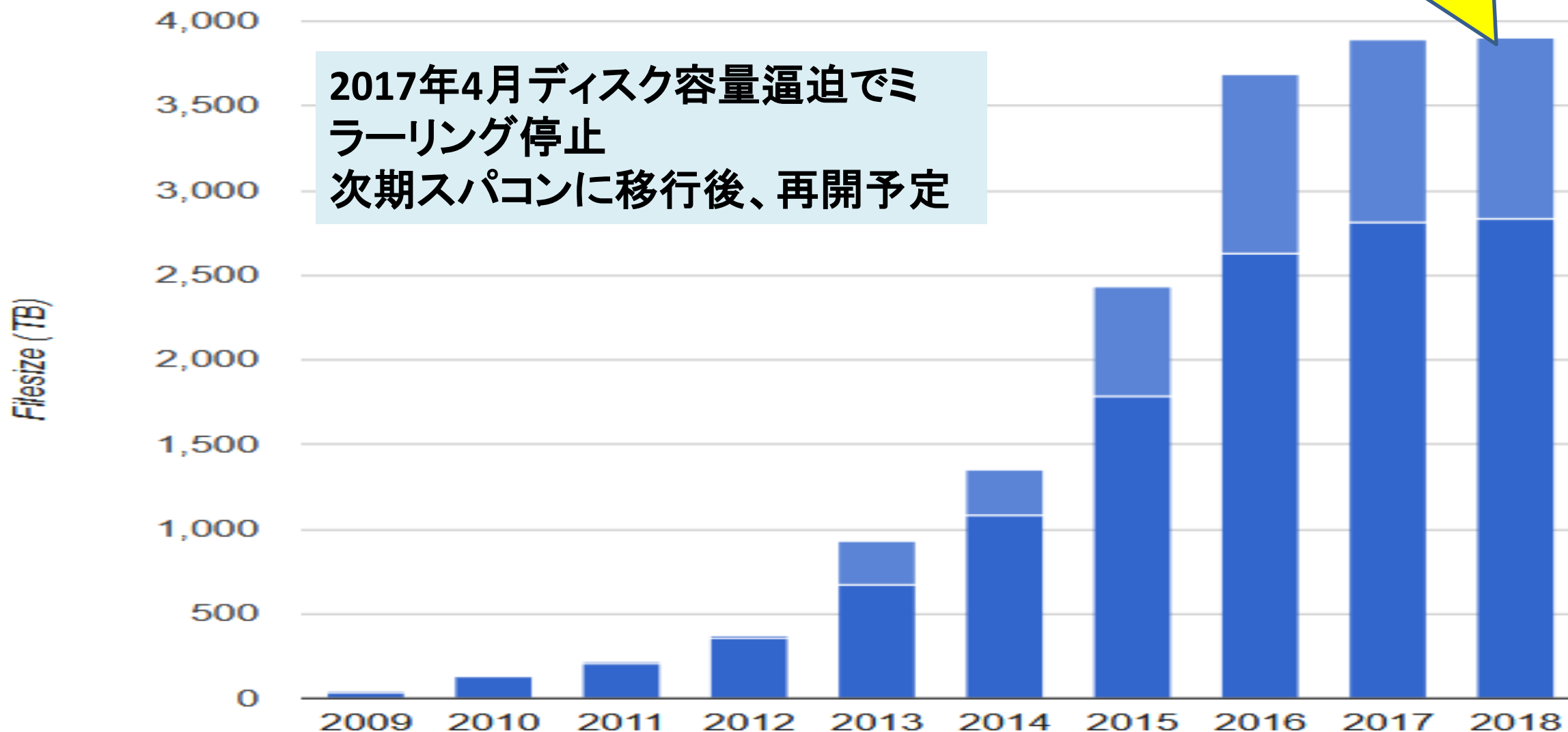
**DRA Run:**

アライメントデータやシーケンスデータを登録

# DRA 公開数

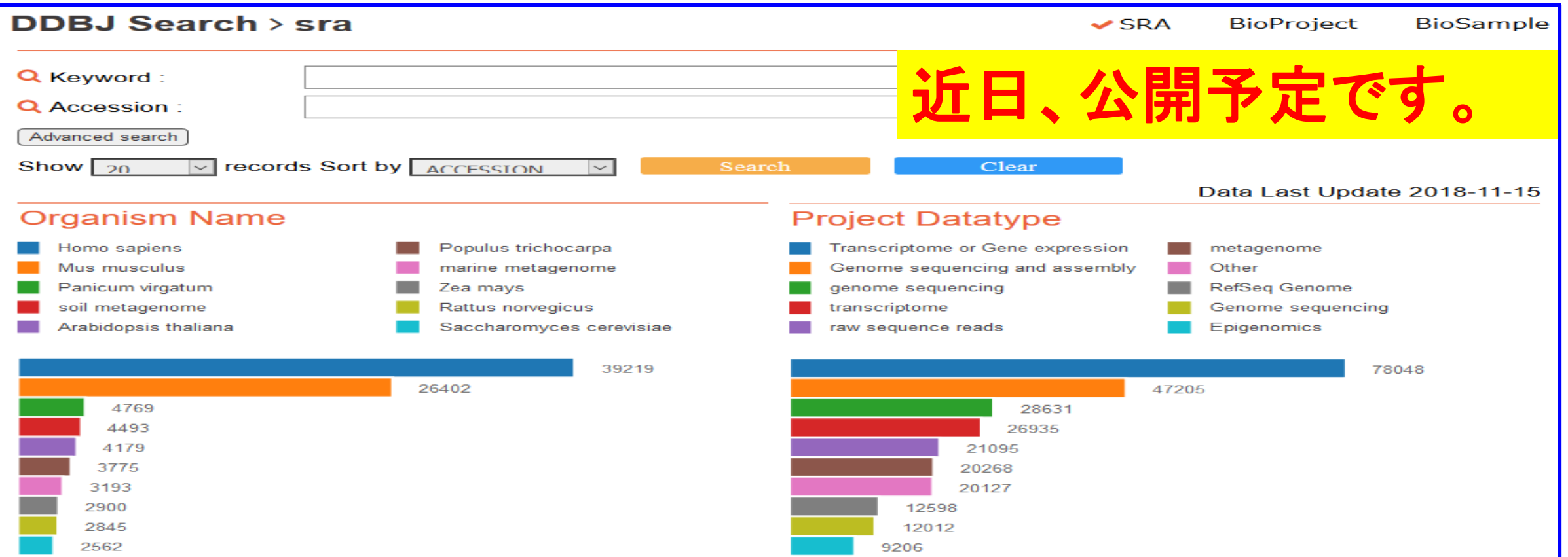
DRA data release (filesize)

■ SRA filesize (TB) ■ fastq filesize (TB)



# NGSデータの新検索ツールの紹介

DBCLSとのコラボレーションにより、NGSデータの新検索ツール(**DDBJ Search**)が開発されました。  
DRA-BioProject-BioSampleを連携して検索可能です。

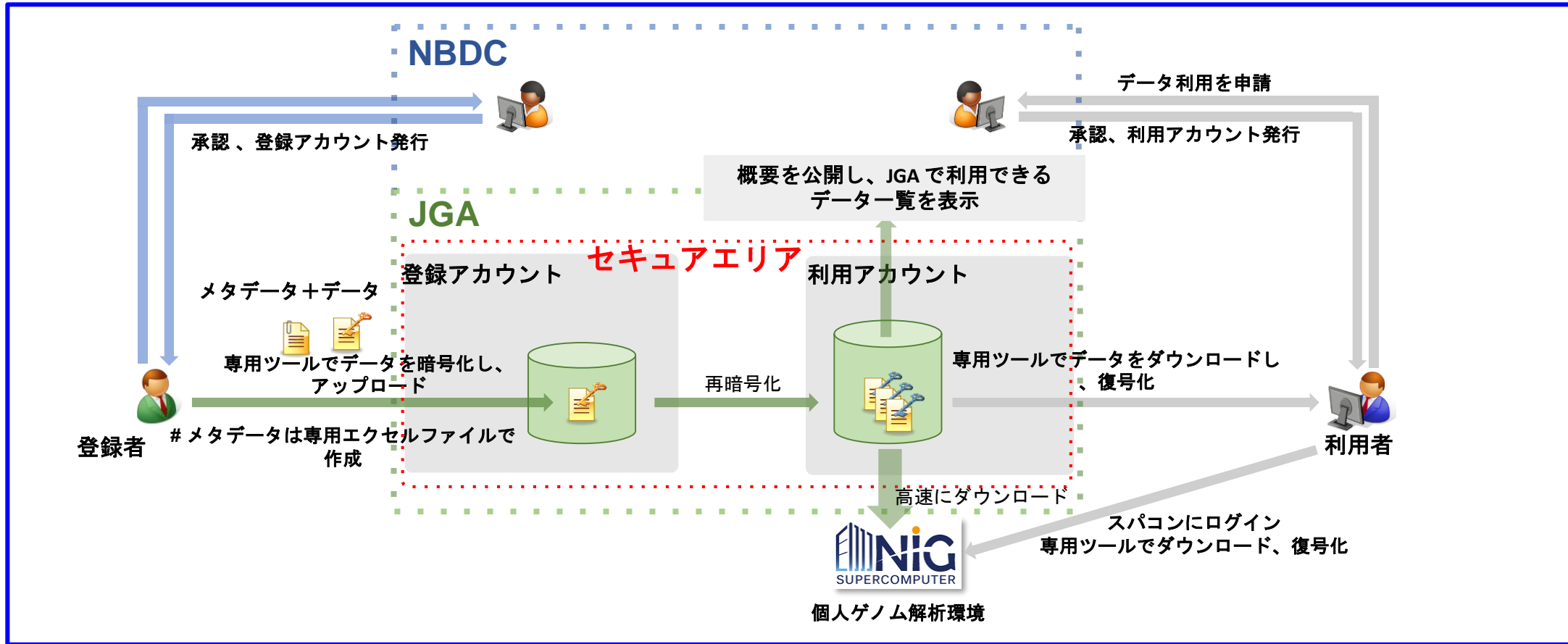


# Japanese Genotype-phenotype Archive (JGA)

JGAは、利用制限が必要なヒト由来の「個人レベルの遺伝学的なデータと匿名化された表現型情報」を保存し提供するアクセス制限データベースです。

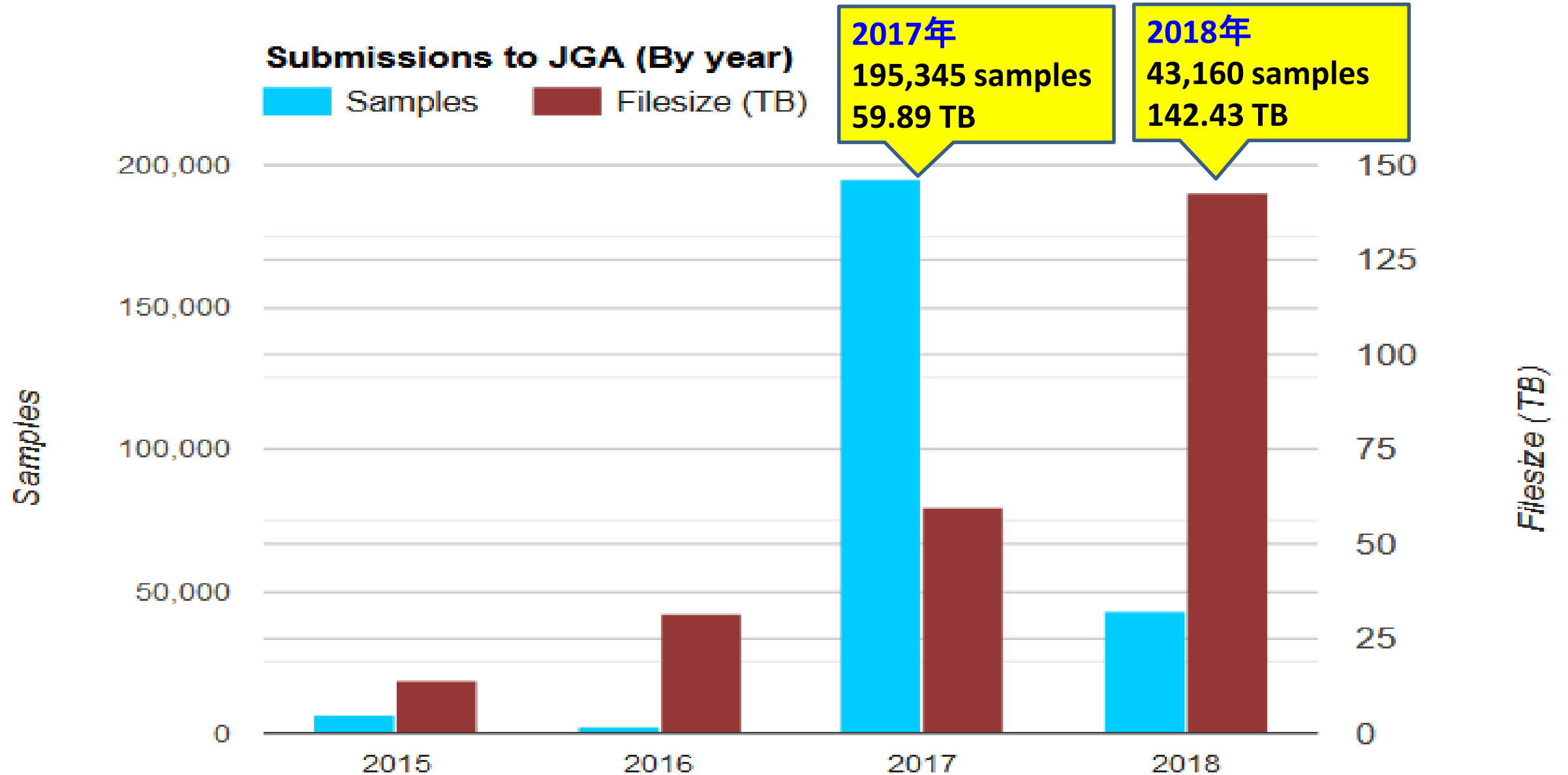
- データが収集された個人との間の同意に基づく協定により、JGA データ利用は**特定の研究目的に制限**されています。  
JGAはセキュアに情報を管理，格納，提供しています。
- JGA は **JST/NBDC ヒトデータ共有ガイドラインに準拠して、NBDCで審査・承認されたデータだけ**を、受け付けています。

# JGA登録・利用フロー



2019年に、NBDC申請、JGA登録利用システムを統合予定です。

# JGA 登録数 (2015-2018)

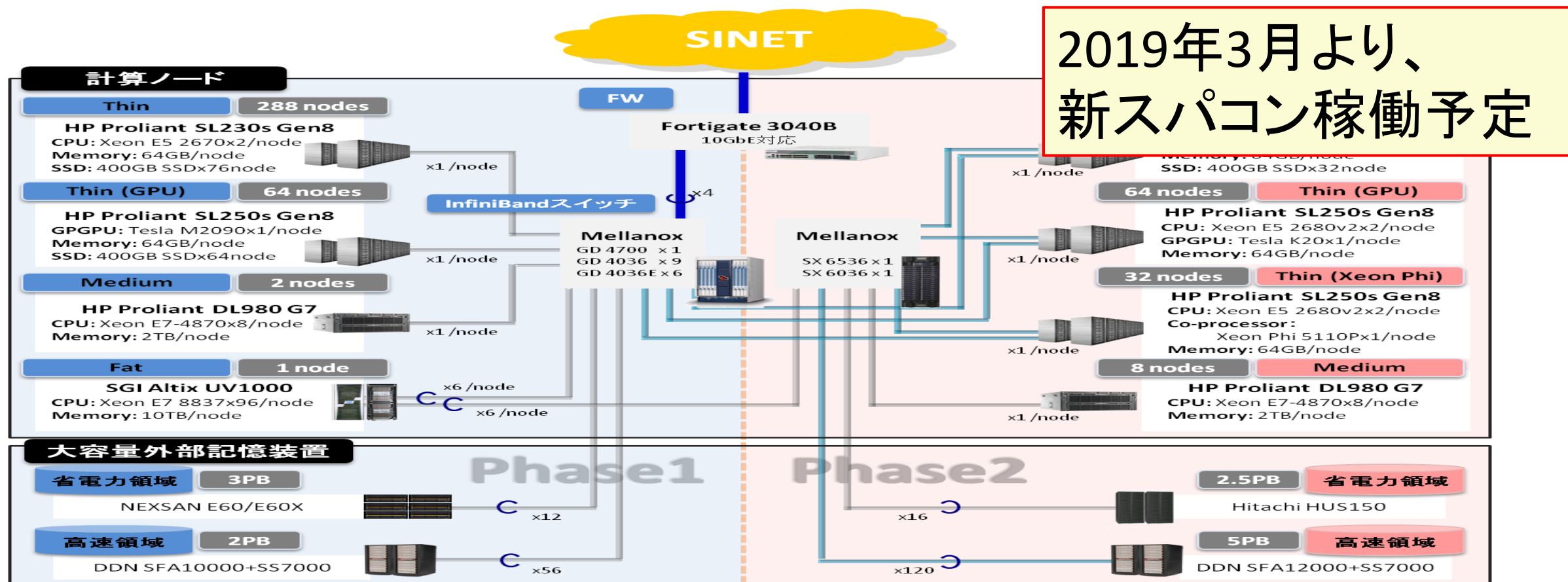


# 第4章

## 国立遺伝学研究所

### スーパーコンピュータシステムの 紹介

# スパコンでデータベースを運用



- スパコンレンタル費: 年間7億円
- 計算用高速ストレージシステム3PB, データベース用大容量ストレージシステム30PBを増強 (2018年3月)



# スパコンサービス

○利用申請をすることで、日本国内の研究者なら、基本的に無料で使用することができます。

○**個人ゲノム解析環境**を用意しています。  
大規模化・複雑化する個人ゲノムデータの解析をサポートするため、2018年9月3日から通常の遺伝研スパコンとは別に、セキュリティを高めた個人ゲノム解析環境の提供を開始しています。

# まとめ

DDBJセンターの使命は、生命情報の共有・解析基盤である一次データベースとスパコンを整備することです。

- 1) データベース構築 と開発
- 2) データ公開・検索サービス  
(定期リリース, getentry, ARSA, DRA Search, BLAST他)
- 3) 国際協力 (GenBank/NCBI, ENA/EBI)
- 4) 他機関との連携協力  
(日本特許庁, 韓国特許庁/KOBIC, NBDC, DBCLS他)
- 5) スパコンの維持とサービス提供

**DDBJセンターは今後も生命科学の発展と共に、  
情報基盤の整備事業を進めていきます。**