

統合データベース講習会：AJACS岐阜  
2013年7月12日

## DDBJを中心とした ゲノムデータベースの紹介

情報・システム研究機構（ROIS） ライフサイエンス統合データベースセンター（DBCLS）  
科学技術振興機構（JST） バイオサイエンスデータベースセンター（NBDC）

河野 信

1

 ©2013 統合データベース講習会 Licensed Under CC 表示 2.1

### 代表的な生命科学関係のデータベース

DNA塩基配列	GenBank/ENA (EMBL)/DDBJ, RefSeq SRA/ENA/DRA
ゲノム	Ensembl, H-InvDB, MGI, FlyBase, TAIR, SGD
タンパク質アミノ酸配列	UniProt (Swiss-Prot + TrEMBL)
タンパク質立体構造	PDB, SCOP, CATH
モチーフ	InterPro, Pfam, PROSITE, ProDom
化合物	PubChem, ChEBI, KEGG LIGAND
パスウェイ	KEGG PATHWAY, Reactome, BioCyc
遺伝子発現	GEO, ArrayExpress, BioGPS
文献	PubMed
その他	GO, NCBI Taxonomy, OMIM, GOLD

3

### 注意点

- ◆ 参加人数が多いため、サイトにつながりにくくなることが予想されます。
- 資料を見ながら適当にタイミングをずらして実行してみてください
- 反応が無くても、何度もクリックしない
  - ますます遅くなるだけです。おおらかな気持ちで臨みましょう
- わからないことがあつたら、講習会のスタッフに気軽に聞いてください

2

 ©2013 統合データベース講習会 Licensed Under CC 表示 2.1

### 本日紹介するDB・ツール

- ◆ DDBJ (DNA Data Bank of Japan)
  - DNA塩基配列を収集
- ◆ Ensembl
  - ゲノムを表示するためのブラウザ
- ◆ BioMart
  - ゲノム情報を取得するためのGUIインターフェイス
- ◆ GOLD (Genome OnLine Database)
  - ゲノムプロジェクトの情報を集めたデータベース

4

 ©2013 統合データベース講習会 Licensed Under CC 表示 2.1

 ©2013 統合データベース講習会 Licensed Under CC 表示 2.1

# DDBJ

DNA Data Bank of Japan

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

5

## DDBJ登録ファイルの例

7

# 国際塩基配列データベースの一覧

◆ International Nucleotide Sequence  
Databank Collaboration (INSDC)

- 米国：GenBank
  - 欧州：ENA
  - 日本：DDBJ



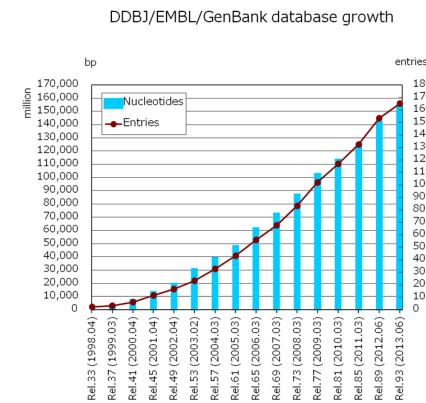
## ◆ (新型) DNAシーケンサーで解読されたDNA塩基配列を収集

©2013 統合云-ターベース講習会 Licensed Under CC 表示 2.1 © 2012 中村保一 (国立遺伝学研究所) licensed under CC表示 2.1日本

## 現在の塩基配列データの量

塩基数：1,500億

登録数：1.6億



Note: CON division is not counted in statistics of DDB1 periodical releases

©2013 統合データベース講習会 Licensed Under CC 表示 2.1 © 2012 中村保一 (国立遺伝学研究所) licensed under CC表示2.1日本

1

## DDBJデータベースを検索してみましょう

9

 ©2012 統合データベース講習会 Licensed Under CC 表示 2.1



DDBJ の紹介 利用の手引き レポート・統計 Q and A お問い合わせ

Google+ カスタム検索 English Search

Web Magazine RSS を購読する DDBJ Twitter

登録 Data Submission 検索・解析 Search / Analysis Super Computer アーカイブ ftp. ddbj.nig.ac.jp

Hot Topics

- 2013.06.26 WABI (Web API for Biology) の再開
- 2013.06.11 DDBJ リリース 93.0, DAD リリース 63.0 完成
- 2013.06.05 DDBJ Web Magazine No.83 発信

Maintenance Information

Last modified : 2013.3.1

## 実習1-1 : DDBJを検索

- ◆DDBJデータベースを”ARSA”という、キーワード検索ツールで検索してみましょう
- 例として大腸菌O157の全ゲノムエントリを検索してみます

DDBJにアクセスするには「DDBJ」で検索  
もしくは <http://www.ddbj.nig.ac.jp/> を直接入力

10

 ©2013 統合データベース講習会 Licensed Under CC 表示 2.1

## DDBJトップページ

11

## DDBJ検索・解析ページ

HOME > 検索・解析

検索・解析

データベース検索

- getentry
- ARSA (アラカルトによるエントリの検索)
- TASERON (高速なキーワード検索)
- blast
- BLAST
- 相同性検索
- DDBJ Vector Screening System
- WABI
- DRA Search

システム解析

- ClustalW
- WABIP

WABIP (Web API for Biology)

WAPI は、DDBJの検索サービスを Web画面 を介して利用できるWeb API です。DDBJ では現在、次のサービスの Web API を提供しています。

WABI BLAST

タングク質データベース及び構造解析

DDBJ Read Annotation Pipeline

WINA

CIB-DDBJ で開発したソフトウェア

Copyright © DNA Data Bank of Japan.

12

 ©2013 統合データベース講習会 Licensed Under CC 表示 2.1

# ARSA

ARSA All-round Retrieval of Sequence and Annotation

E. coli O157 を検索

QuickSearch E. coli O157 Search

検索条件を複数入力する場合は、&(AND条件)、|(OR条件)、!(AND NOT条件)を指定することが可能です。

ARSA All-round Retrieval of Sequence and Annotation

QuickSearch E. coli O157 Search

検索条件を複数入力する場合は、&(AND条件)、|(OR条件)、!(AND NOT条件)を指定することが可能です。

DB(2195) Patient\_A(M3)

PrimaryAccessionNumber Definition moltype Organism Length

AB011548	Escherichia coli O157:H7 str. Sakai plasmid pOSAK1 DNA, complete sequence.	DNA	Escherichia coli O157:H7 str. Sakai	3306
AB011549	Escherichia coli O157:H7 str. Sakai plasmid pO157 DNA, complete sequence.	DNA	Escherichia coli O157:H7 str. Sakai	92721
AB035920	Escherichia coli O157:H7 hemG, rrsA, ileT, alaT, rrlA, rrfA, mobB, mobA genes for protoporphyrin oxidase protein, 16S rRNA, isoleucine tRNA 1, alanine tRNA 1B, 23S rRNA, 5S rRNA, molybdopterin-guanine dinucleotide biosynthesis protein B, molybdopterin-guanine dinucleotide biosynthesis protein A, complete and partial cds.	DNA	Escherichia coli O157:H7	7003

検索結果：  
ヒット数が多くるので  
絞り込みが必要  
詳細検索  
(Advanced Search)へ

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

13

14

ヒットしない…

ARSA All-round Retrieval of Sequence and Annotation

Refine Search

DB(5)

All Select

PrimaryAccessionNumber Definition moltype Organism Length

AB602479	C. glutamicum-E. coli shuttle vector pCRB12 DNA, complete sequence.	DNA	C. glutamicum-E. coli shuttle vector pCRB12	4569
AB671168	E. coli-T. thermophilus shuttle vector pTRK1T DNA, complete sequence.	DNA	E. coli-T. thermophilus shuttle vector pTRK1T	7482
AB671169	E. coli-T. thermophilus shuttle vector pTRH1T DNA, complete sequence.	DNA	E. coli-T. thermophilus shuttle vector pTRH1T	6057
HM126493	C. glutamicum-E. coli shuttle vector pCRB62, complete sequence.	DNA	C. glutamicum-E. coli shuttle vector pCRB62	5914
HM126494	C. glutamicum-E. coli shuttle vector pCRB12, complete sequence.	DNA	C. glutamicum-E. coli shuttle vector pCRB12	4569

[1]

原因：Organismの項目には正式名称しか書かれていないため

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

15

# いくつかの特徴で絞り込み

ARSA All-round Retrieval of Sequence and Annotation

DB(Quick Search) DB(Advanced Search)

- フィールド内で検索条件を複数入力する場合は、& (AND条件)、| (OR条件)、! (AND NOT条件)を指定することができます。
- ダブルクォーテーション("")で囲まれた文字列は、1つのキーワードとして認識されます。
- 検索方法および検索条件の例などを知りたい方は[こちら](#)をクリックして下さい。

Search reset

Combine Searches with  & (AND)

All Text

Accession Number

Primary Accession Number

Division  BCTC  CCON  ENV  HTC  HTG  HUM  INV

MAM  PAT  PHG  PLN  PRI  ROD  STS

SYN  TSIA  UNA  VRL  VRV

Sequence Length

Molecular

Type  DNA  RNA  cRNA  mRNA  rRNA  tRNA

Form  circular  linear

Date

Definition

Comment

Keyword

Organism  E. coli

Taxon

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

14

# Escherichia coli O157 で再検索

Molecular

Type  DNA  RNA  cRNA  mRNA  rRNA  tRNA  
Form  circular  linear

Date

Definition

Comment

Keyword

Organism  Escherichia coli O157

Taxon

ARSA All-round Retrieval of Sequence and Annotation

DB(12) DB(Quick Search) DB(Advanced Search)

Refine Search

BA000007 をクリックすると  
O157のゲノムエントリを表示

チェックを入れて"Download"を  
クリックするとエントリ、  
配列をダウンロード可能

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

16

PrimaryAccessionNumber Definition moltype Organism Length

AB011548	Escherichia coli O157:H7 str. Sakai plasmid pOSAK1 DNA, complete sequence.	DNA	Escherichia coli O157:H7 str. Sakai	3306
AB011549	Escherichia coli O157:H7 str. Sakai plasmid pO157 DNA, complete sequence.	DNA	Escherichia coli O157:H7 str. Sakai	92721
AEO05174	Escherichia coli O157:H7 EDL933, complete genome.	DNA	Escherichia coli O157:H7 str. EDL933	5528445
AF074613	Escherichia coli O157:H7 str. EDL933 plasmid pO157, complete sequence.	DNA	Escherichia coli O157:H7 str. EDL933	92077
BCI(11) CMI(1)	Escherichia coli O157:H7 str. Sakai DNA, complete genome.	DNA	Escherichia coli O157:H7 str. Sakai	5498450
CM000662	Escherichia coli O157:H7 str. TW14588 chromosome, whole genome shotgun sequence.	DNA	Escherichia coli O157:H7 str. TW14588	5578816
CP001163	Escherichia coli O157:H7 str. EC4115 plasmid pO157, complete sequence.	DNA	Escherichia coli O157:H7 str. EC4115	94644
CP001164	Escherichia coli O157:H7 str. EC4115, complete genome.	DNA	Escherichia coli O157:H7 str. EC4115	55716
ECI(15)	Escherichia coli O157:H7 str. EC4115 plasmid pEC4115, complete sequence.	DNA	Escherichia coli O157:H7 str. EC4115	37452

## 実習1-2 : NCBIを検索

◆先ほどと同じ検索をNCBI（GenBankの提供機関）の検索システム"GQuery"で実行してみましょう

○例として大腸菌O157の全ゲノムエントリを検索してみます

NCBI GQueryにアクセスするには「gquery」で検索もしくは <http://www.ncbi.nlm.nih.gov/sites/gquery> を直接入力

17

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

## NCBI Entrez Search

The screenshot shows the NCBI Entrez search interface. At the top, there's a navigation bar with links for HOME, SEARCH, SITE MAP, PubMed, All Databases, Human Genome, GenBank, Map Viewer, and BLAST. Below the navigation bar is a search bar labeled "Search across databases" with a "GO" button and a "Clear" button. To the right of the search bar is a help link. The main area is titled "Welcome to the Entrez cross-database search page". It lists several databases with icons: PubMed (biomedical literature citations and abstracts), PubMed Central (free, full-text journal articles), Site Search (NCBI web and FTP sites), Nucleotide (core subset of nucleotide sequence records), EST (Expressed Sequence Tag records), GSS (Genome Survey Sequence records), Protein (sequence database), Genome (whole genome sequences), Structure (three-dimensional macromolecular structures), Taxonomy (organisms in GenBank), and SNP (short genetic variations). On the left, there are filter boxes for 文献 (Literature), 塩基配列 (Nucleotide Sequences), アミノ酸配列 (Protein), ゲノム配列 (Genomes), and 立体構造 (Structure). At the bottom left is a copyright notice: "©2013 統合データベース講習会 Licensed Under CC 表示 2.1".

## GQuery: E. coli O157 で検索

The screenshot shows the GQuery search results for "E. coli O157". The search term "E. coli O157" is highlighted in a red oval. The results are categorized into sections: Literature, Health, Organisms, Nucleotide Sequences, and Genomes. In the "Genomes" section, the "E. coli O157" entry is also highlighted in a red oval. Other entries in this section include "Genome sequencing projects by organism" (2), "Assembly: genomic assembly information" (0), "Epigenomics: epigenomic studies and display tools" (102), and "BioSample: descriptions of biological source materials" (138). At the bottom left is a copyright notice: "©2013 統合データベース講習会 Licensed Under CC 表示 2.1".

19

## Entrez: 大腸菌ゲノムページ

The screenshot shows the NCBI Entrez genome page for "Escherichia coli". The search term "Escherichia coli" is highlighted in a red oval. The page includes sections for "Organism Overview", "Representatives", and "Dendrogram (based on genomic BLAST)". The "Dendrogram" shows the phylogenetic relationship between various Escherichia coli strains, with "Escherichia coli O157:H7 str. Sakai" highlighted in a red oval. Other strains shown include DH1, BL21(DE3), B str. REL606, ETEC H10407, O104:H4 str. 2009EL-2071, O104:H4 str. 2011C-3493, O104:H4 str. 2009EL-2050, 55989, O157:H7 str. EC4115, O157:H7 str. W1359, O157:H7 str. EDL933, and O157:H7 str. Sakai. At the bottom left is a copyright notice: "©2013 統合データベース講習会 Licensed Under CC 表示 2.1".

20

## 大腸菌O157 Sakai株のページ

Display Settings: Overview  
Send to: Related information  
BioProject  
Gene  
Components  
Protein  
PubMed  
Taxonomy

Recent activity  
Turn Off Clear

Representative  
Reference genome, Community selected: Escherichia coli O157:H7 str. Sakai ASM886v1

Genome Sequencing Projects

Organism	BioProject	Assembly	Status	Chrs	Plasmids	Size (Mb)	GC%	Gene	Protein
Escherichia coli O157:H7 str. Sakai	PRJNA57781, PRJNA226	ASM886v1	●	1	2	5.59	50.4	5,460	5,318

See more... Go to nucleotide Graphics FASTA GenBank

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

21

## 配列を取得したとの解析例

- ◆ “blast”等の配列類似性検索を実行して、類似の配列を収集する
- ◆ “primer3”等で配列をクローニングするためのプライマーを設計する
- ◆ “clustalW”等でマルチプルアラインメントを作成し、配列の共通部分や進化関係を調べる
- ◆ “interproscan”等でモチーフ構造を調べる
- ◆ “swiss-model”等で立体構造を予測する

23

## 豊富なリンクとツール群

Display Settings: GenBank  
Sequence not displayed. Use 'Customize View' section for control.

Escherichia coli O157:H7 str. Sakai chromosome, complete genome  
NCBI Reference Sequence: NC\_002695.1  
FASTA Graphics

Go to: Locus NC\_002695 5498450 bp DNA circular BCT 25-JAN-2012  
DEFINITION Escherichia coli O157:H7 str. Sakai chromosome, complete genome.  
ACCESSION NC\_002695  
VERSION NC\_002695.1 Gt15829254  
DBLINK Project: 57781  
KEYWORDS SOURCE: Escherichia coli O157:H7 str. Sakai  
ORGANISM Escherichia coli O157:H7 str. Sakai  
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;  
Enterobacteriaceae; Escherichia.  
REFERENCE 1 (bases 1 to 5498450)  
AUTHORS Bergholz,T.M., Wick,L.M., Q,W., Riordan,J.T., Ouellette,L.M. and  
Whitam,T.S.  
TITLE Growth and nutritional response of Escherichia coli O157:H7 to  
growth transitions in glucose minimal medium  
JOURNAL BMC Microbiol. 7, 57 (2007)  
PUBMED 17967175  
REMARK Publication Status: Online-Only  
REFERENCE 2 (sites)  
AUTHORS Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Itoh,K.,  
Yokota,K., Han,C.G., Otsubo,E., Nakayama,K., Murata,T.,  
Tanaka,M., Tobe,T., Iida,T., Takami,H., Honda,T., Sasakawa,C.,  
Ogasawara,N., Yasunaga,T., Kuhara,S., Shiba,T., Hattori,M. and  
Shinagawa,H.  
TITLE Complete genome sequence of enterohemorrhagic Escherichia coli  
O157:H7 and genomic comparison with a laboratory strain K-12  
JOURNAL DNA Res. 8 (1), 11-22 (2001)  
PUBMED 11753798  
REMARK Erratum:[DNA Res 2001 Apr 27;8(2):96]

BLASTによる類似性検索  
プライマー設計ツール  
文献へのリンク

LinkOut to external resources  
REBASE enzyme XbaMNP [REBASE - The Restriction Enzy...]  
REBASE enzyme M.EcoP93Damp [REBASE - The Restriction Enzy...]  
REBASE enzyme M.EphHK97Damp [REBASE - The Restriction Enzy...]  
REBASE enzyme M.EcoCR3Fp [REBASE - The Restriction Enzy...]  
REBASE enzyme M.EcoV2Dam [REBASE - The Restriction Enzy...]  
REBASE enzyme S.EcoKO157ORFAP [REBASE - The Restriction Enzy...]  
REBASE enzyme EcoKO157ORF5262P [REBASE - The Restriction Enzy...]

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

22

## 新型シーケンサからのデータ

©2012 統合データベース講習会 Licensed Under CC 表示 2.1

24

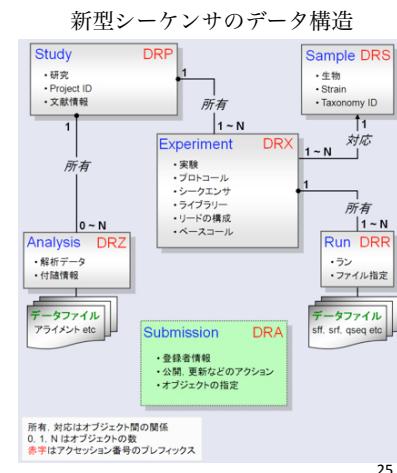
DDBJ Sequence Read Archive (DRA)

#### ◆新型シーケンサデータを保存・共有



登録されているデータ構造は少々複雑ですが、  
DRAのページでは「日本語」での詳しい説明がある  
登録されているデータはあまりうまく  
整理されていない

 ©2013 統合データベース講習会 Licensed Under CC 表示 2.0



## SRAs: Survey of Read Archives

◆統計値から、分類をたどってデータにアクセス  
することも可能

### Search by statistics - 統計値から探す

The number of projects are indicated in "Study Types" table. The totals in "Platforms" and "Species of Samples" are larger than one of "Study Types" because a project can contain some platforms and samples.

実際のプロジェクト数は"Study Types"に書かれているものです。"Platforms"や"Species of Samples"のtotalの値は、複数のプラットフォームで行われた実験が1つのプロジェクトでなされる場合が多々あります。ダブルカウントしているので、数字が大きくなっています。

2012-06-21 updated

#### **Study Types**

**Whole Genome Sequencing**

**Transcriptome Analysis**

**Metagenomics**

**Epigenetics**

**Resequencing**

**Other**

**RNASeq**

**Population Genomics**

**Gene Regulation Study**

**Cancer Genomics**

**Exome Sequencing**

**Pooled Clone Sequencing**

**Synthetic Genomics**

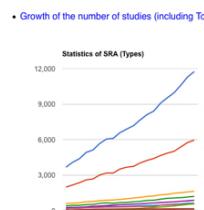
**Forensic or Paleo-genomics**

Total

12008

• Growth of the number of studies (including Total)

Statistics of SRA (Types)



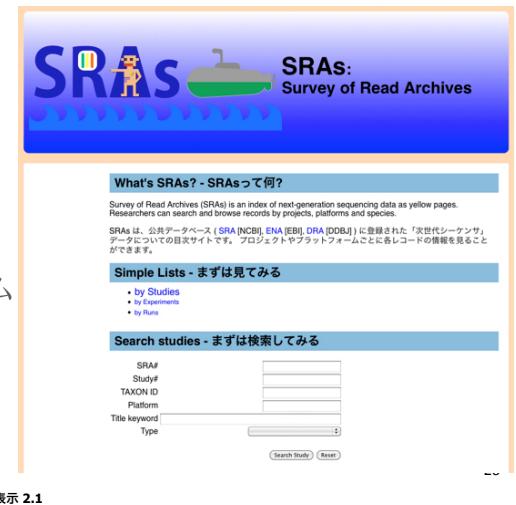
- Growth of the number of studies (without Total)

?

## SRAs: Survey of Read Archives

## ◆SRA/DRAに登録されているデータを メタデータで整理

◆ <http://sra.dbcls.jp/>



## ●生物種

## ○解析プラットフォーム

## ○キーワード

などで検索可能

鎖鋸 (kusarinoko)

#### ◆論文が出ているSRA/DRAエントリのまとめ

- 論文が出ているということは、査読を経ているので、一定のデータの質は担保されている（はず）

◆<http://g86.dbcls.jp/kusarinoko>

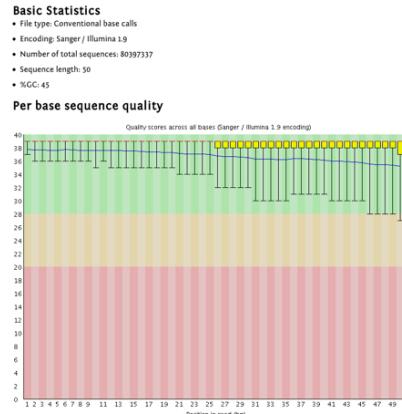


(CC) BY ©2013 総合データベース講習会 | licensed Under CC 表示 2.1

28

## 鎖鋸 (kusarinoko)

- ◆独自に”FastQC”をかけて、  
それぞれのデータの質を評価



©2013 統合データベース講習会 Licensed Under CC 表示 2.1

29

# Ensembl

## Ensembl Genome Browser

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

31

## DDBJ Read Annotation Pipeline

- ◆新型シーケンサデータの解析パイプライン

○<http://p.ddbj.nig.ac.jp/>

The screenshot shows the "Selecting Tools for Basic Analysis of DDBJ ANNOTATION PIPELINE" page. It includes sections for "Reference Genome Mapping" (listing tools like BLAT, Mezzi, bwa, SCALe, Bowtie, and Trimmomatic) and "do novo Assembly" (listing tools like SCALe, AbusS2, Velvet, and Trinity). Below these are tabs for "Job Confirmation", "Preprocessing status", and "Mapping status". A sidebar on the left provides links to DDBJ Read Annotations, the Development Team, and other resources.

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

30

## Ensembl

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

- ◆Wellcome Trust Sanger Instituteと  
European Bioinformatics Institute (EBI)が  
共同開発しているゲノムブラウザ

○さまざまなアノテーション情報をゲノム上で見る  
(ブラウズ) することが可能

○脊椎動物を中心

○EnsemblBacteria, EnsemblFungi, EnsemblMetazoa,  
EnsemblPlants, EnsemblProtistsというのもある

- ◆他にUCSC Genome Browserが有名

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

32

## 実習2：Ensemblを使ってみる

### ◆ヒトアセトアルデヒドデヒドロゲナーゼ2 (ALDH2) 周辺のアノテーション情報を眺める

生物種を選択して ALDH2 で検索

The screenshot shows the Ensembl homepage with the search bar set to "Human". The genome browser on the right displays the ALDH2 gene on chromosome 12. A red box highlights the search term "ALDH2" in the search bar.

33

### サマリページの表示

ゲノムブラウザで表示

The screenshot shows the Ensembl gene summary page for ALDH2. It includes sections for Gene summary, Alternative genes, and a detailed genomic track viewer. A red box highlights the "Region in detail" button in the sidebar.

35

## 対象遺伝子の選択

The screenshot shows the Ensembl results summary page for the query "ALDH2". It lists one result for "Human (H1)". A red box highlights the "Gene" section in the results table.

ALDH2 (H1) - Human (GRCh37)

Gene ID: ENSG00000111275

Description: aldehyde dehydrogenase 2 family (mitochondrial) [Source:HGNC Symbol;Acc:404] (Type: protein coding Ensembl/Havana merge)

Gene ID: ENSG00000111275

Location: 12:112,204,691-112,247,782

Variations: 27

Source: e72

34

### ゲノムブラウザでの表示

The screenshot shows the Ensembl genome browser for the ALDH2 gene on chromosome 12. It displays genomic tracks for chromosomes 12 and 13, with a red box highlighting the "Configure this page" button in the sidebar.

Chromosome 12: 112,204,691-112,247,782

Region in detail

Configure this page

36

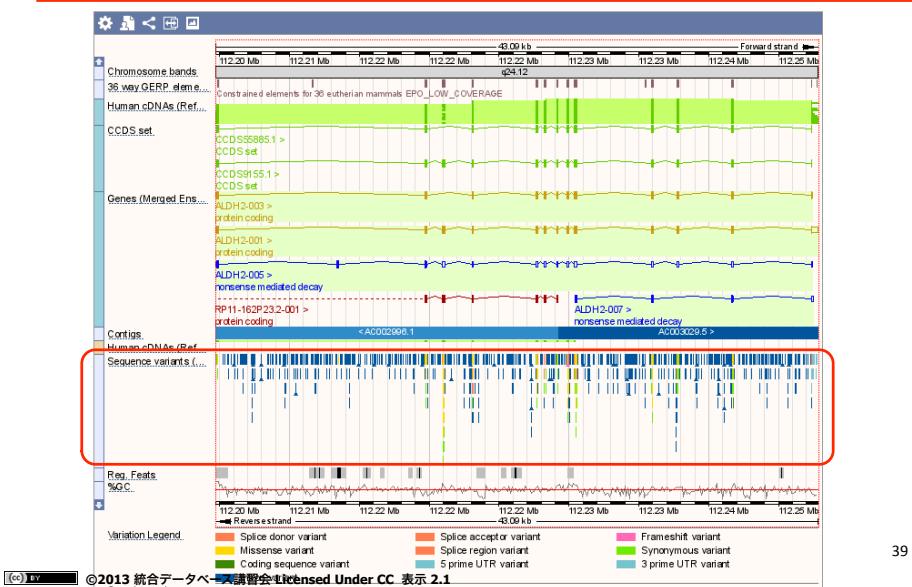
## 表示するアノテーションの指定

表示させる項目を選択

The screenshot shows the 'Active tracks' section of the Ensembl genome browser. On the left, a sidebar lists various categories like 'Sequence and assembly', 'Genes and transcripts', and 'Regulation'. On the right, a track list shows items such as 'Constrained elements for 36 eutherian mammals EPO\_LOW\_COVERAGE' and 'Reg. Feats'. A red box highlights the 'Active tracks' tab in the top navigation bar.

37

## 選択したアノテーション情報が表示される



39

## 表示するアノテーションの指定

設定後、✓マークをクリックすると自動更新される

The screenshot shows the 'Variation' configuration page. It includes sections for 'Sequence and assembly' (with options like 'Collapsed' and 'Expanded'), 'mRNA and protein alignments', and 'ncRNA'. A red box highlights the 'Variation' section in the main list. A red checkmark is visible in the top right corner of the window.

38

## 自分のデータをゲノムブラウザ上に表示

This screenshot shows the 'Add your data' section of the configuration interface. It includes fields for 'Name for this data (optional)', 'Species' (set to 'Human (Homo sapiens)'), and 'Data format' (with a dropdown menu showing 'BED'). A red box highlights the 'Add your data' button. A red checkmark is visible in the top right corner of the window.

40

既存のアノテーションと  
自分のデータをゲノム上で  
重ね合わせて表示することが可能

# BioMart

41

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

## BioMart



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

- ◆ Ontario Institute for Cancer Research (OICR)と European Bioinformatics Institute (EBI)が共同開発しているデータ管理システム
- ◆ さまざまなデータベースのGUIデータ取得インターフェイスとして利用されている
- Ensembl, UniProt, WormBase, KazusaMartなど

42

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

### 実習3：BioMartを使ってEnsemblからデータ取得

- ◆ ヒトゲノムの22番染色体にある遺伝子について、転写開始点から上流の配列1kbをFASTA形式で取得する

Human (GRCh37) ▾ 12:112,204,691-112,247,782 Gene: ALDH2

Location-based displays

- Whole genome
- Chromosome summary
- Region view
- Region comparison
- Region in detail
- Comparative Genomics
- Alignments (image) (64)
- Alignments (text) (64)
- Region Comparison (69)
- Comparative Genomics (7)
- Genetic Variation
- Resequencing (20)
- Linkage Data
- Markers
- Other genome browsers
- UCSC
- NCBI
- Vega

Configure this page Add your data Export data Bookmark this page

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

43

### データセットを選択

Ensembl beta

BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors | Search Human... Search

New Count Results URL XML Perl Help

Dataset [None selected]

✓ - CHOOSE DATABASE -

- Ensembl Genes 72
- Ensembl Variation 72
- Ensembl Regulation 72
- Vega 52
- PRIDE (EBI UK)

New Count Results URL XML Perl Help

Dataset [None selected]

✓ - CHOOSE DATASET -

- Danio rerio genes (Zv9)
- Gallus gallus genes (GalGal4)
- Homo sapiens genes (GRCh37.p11)
- Mus musculus genes (GRCm38.p1)
- Rattus norvegicus genes (Rnor\_5.0)

Aluropoda melanoleuca genes (allMe1)

Anolis carolinensis genes (AnoCar2.0)

Bos taurus genes (UMD3.1)

Caenorhabditis elegans genes (WBcel235)

Callithrix jacchus genes (calJac3)

Canis familiaris genes (CanFam3.1)

Cavia porcellus genes (cavPor3)

Chlorocebus aethiops genes (chtoHot1)

Ciona intestinalis genes (CIRN1)

Ciona savignyi genes (CSAV2.0)

Dasyurus maculatus genes (daxNov2)

Dipodomys ordii genes (dipOrd1)

Drosophila melanogaster genes (BDGP5)

Echinops telfairii genes (TENREC)

Equus caballus genes (EquCab2)

Erinaceus europaeus genes (erlEur1)

Felis catus genes (Felis\_catus\_6.2)

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

## 取得したい遺伝子セットを指定

The screenshot shows the Ensembl BioMart search interface. In the top navigation bar, 'Dataset' is selected under 'Homo sapiens genes (GRCh37.p11)'. The 'Filters' section is highlighted with a red box. Below it, the 'Attributes' section is also highlighted with a red box. The main search area contains a form with various filters like 'REGION', 'GENE', 'TRANSCRIPT EVENT', etc., all of which are circled in red.

Dataset  
Homo sapiens genes (GRCh37.p11)  
Filters  
[None selected]  
Attributes  
Ensembl Gene ID  
Ensembl Transcript ID  
Dataset  
[None Selected]

Please restrict your query using criteria below

REGION:  
GENE:  
TRANSCRIPT EVENT:  
GENE ONTOLOGY:  
EXPRESSION:  
MULTI SPECIES COMPARISONS:  
PROTEIN DOMAINS:  
VARIATION:

45

## 取得したいデータの種類を指定

The screenshot shows the Ensembl BioMart search interface. In the top navigation bar, 'Dataset' is selected under 'Homo sapiens genes (GRCh37.p11)'. The 'Attributes' section is highlighted with a red box. Below it, the 'Filters' section is also highlighted with a red box. The main search area contains a form with various attributes like 'Features', 'Structures', 'Variation', etc., all of which are circled in red.

Dataset  
Homo sapiens genes (GRCh37.p11)  
Filters  
Chromosome: 22  
Attributes  
Ensembl Gene ID  
Ensembl Transcript ID  
Dataset  
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

Features      Homologs  
Structures      Variation  
Transcript Event      Sequences  
GENE:  
EXTERNAL:  
EXPRESSION:  
PROTEIN DOMAINS:

47

## 取得したい遺伝子セットを指定

The screenshot shows the Ensembl BioMart search interface. In the top navigation bar, 'Dataset' is selected under 'Homo sapiens genes (GRCh37.p11)'. The 'Attributes' section is highlighted with a red box. The main search area contains a form with a 'Chromosome' dropdown set to '22' and a 'Base pair' input field with '1' and '10000000'.

Dataset 1209 / 63253  
Genes  
REGION:  
Chromosome: 22  
Attributes  
Ensembl Gene ID  
Flank (Gene)  
Upstream flank [1000]  
Associated Gene Name  
Dataset  
[None Selected]

Please restrict your query using criteria below

Base pair  
Gene Start (bp)  
Gene End (bp)

46

## 取得したいデータの種類を指定

The screenshot shows the Ensembl BioMart search interface. In the top navigation bar, 'Dataset' is selected under 'Homo sapiens genes (GRCh37.p11)'. The 'Attributes' section is highlighted with a red box. The main search area contains a form with a 'Flank (Gene)' dropdown selected and an 'Upstream flank' input field set to '1000'.

Dataset  
Homo sapiens genes (GRCh37.p11)  
Filters  
Chromosome: 22  
Attributes  
Ensembl Gene ID  
Ensembl Transcript ID  
Flank (Gene)  
Upstream flank []  
Dataset  
[None Selected]

SEQUENCES:  
Sequences (max 1)

Unspliced (Transcript)  
Unspliced (Gene)  
Flank (Transcript)  
Flank (Gene)  
Flank-coding region (Transcript)  
Flank-coding region (Gene)

Upstream flank  
Upstream flank [1000]

Downstream flank  
Downstream flank []

Header Information

48

取得したいデータの種類を指定

ヘッダ情報（FASTAのコメント行）の指定

 **Ensembl** GENOME  
[BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#)

[Login/Register](#)

🔍

---

[New](#) [Count](#) [Results](#)

 URL  XML  Perl  Help

---

**Dataset**

Homo sapiens genes  
(GRCh37.p11)

**Filters**

Chromosome: 2,22

**Attributes**

Ensembl Gene ID  
Flank (Gene)  
Upstream flank [1000]  
Associated Gene Name

---

**Dataset**

[None Selected]

**Header Information**

**Gene Information**

- Ensembl Gene ID
- Description
- Associated Gene Name
- Associated Gene DB
- Chromosome Name

---

**Transcript Information**

- CDS start (within cDNA)
- CDS end (within cDNA)
- 5' UTR Start
- 5' UTR End
- 3' UTR Start
- 3' UTR End

---

**Exon Information**

- CDS Length
- CDS Start

Resultsで取得データを表示

4

©2013 統合データベース講習会 Licensed Under CC 表示 2.1

## データのダウンロード

50

# GOLD

## Genome OnLine Database

5

# GOLD: Genome OnLine Database

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

- ◆ 1995年の*Haemophilus influenzae*ゲノムの解読以来、現在までにゲノムが解読された／解読中の生物を集めたDB
  - ◆ 日本で最初に解読されたゲノムは光合成細菌である  
*Synechocystis*のゲノムで、1996年（かずさDNA研究所）
  - ◆ 真核生物最初のゲノムは1997年の*Saccharomyces cerevisiae*（出芽酵母）
  - ◆ メタゲノムプロジェクト（ある環境にいる微生物のゲノムをまとめて読む）もある
    - 腸内細菌、皮膚常在菌、海水、排水、etc...

52

## 実習4-1：ゲノム解読済みの生物を眺める

**GOLD**  
Genomes Online Database

Last update: 2013-07-08  
Total # of genomes: 27099  
Download GOLD

[Home](#)

Genome Map  
Genome Earth  
Search  
News  
Statistics  
Team  
Reference  
Contact

[Classification](#)  
• Studies: 392  
• Samples: 0

**Welcome to the Genomes OnLine Database**

GOLD: Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world.

**Metagenomes**

**Isolate Genomes**

**Genome Distribution**

• Project Type  
• Sequencing Status  
• Phylogenetic

**1. Register**  
**GOLD**  
Genomes Online Database  
Register your project information and Metadata in Genomes Online Database  
**Register**

**2. Annotate**  
**SIGS**  
Standards in Generic Sciences  
An Open Access Journal of the Generic Standards Consortium  
Annotate your microbial genome or metagenome with IMG/ER or IMG/MER  
**Annotate**

**3. Publish**  
**SIGS**  
Standards in Generic Sciences  
An Open Access Journal of the Generic Standards Consortium  
Publish your genome or metagenome in open access standards-supportive journal.  
**Publish**

U.S. DEPARTMENT OF ENERGY Office of Science

©2012 The Regents of the University of California  
Disclaimer | Credits

Version 4.0

## 実習4-1：ゲノム解読済みの生物を眺める

ID	生物種名	分類	ゲノムサイズ	データ	データベース	終了日					
G000001	Escherichia coli K-12, MG1655	B	PROTEOBACTERIA-GAMMA Escherichia Enterobacteriaceae Escherichia Wikipedia	4640 Kb 4449 4749 4942 5042	1	51% U00096 U00096	Univ of Wisconsin-Madison	Cellulose Ecology Ecogenomics IMB-Annotation I-Craig Venter Institute NCBI NCBI NCBI	Science 272(5313):1174 1997-09-05	1997-09-05	Bateman R
G000002	Helicobacter pylori 26695	B	PROTEOBACTERIA-EPILSON Helicobacter Enterobacteriaceae Helicobacter Wikipedia	1885 Kb 1624 1745 1857 1957	1	39% AE000511 AE000511	I-Craig Venter Institute	IBM-Annotation I-Craig Venter Institute NCBI NCBI NCBI	Nature 338(6153):547 1997-08-07	1997-08-07	Tomb JF
G000006	Saccharomyces cerevisiae S288C	E	FUNGI-ASCOMYCOTA Saccharomyces Enterobacteriaceae Saccharomyces Wikipedia	12157 Kb 8273 8723 9176 9527	16	38% AB011856 C00000000 K29720	International Collaboration	GenoDB Geogaud IMB-Annotation I-Craig Venter Institute NCBI NCBI NCBI NCBI NCBI NCBI	Nature 387:5-10 1997-05-29	1997-05-29	
G000005	Mycoplasma pneumoniae M129	B	Tenericutes Tenericutes Enterobacteriaceae Mycoplasma Wikipedia	816 Kb 733 760 804 840	1	40% U00089	Univ of Heidelberg	IBM-Annotation I-Craig Venter Institute Holmgren NCBI ZBH	Nucleic Acids Research 24(4):4449 1999-11-15	1996-11-15	Herrmann R
G000004	Methanococcoides jannaschii DSM 2661	A	Euryarchaeota-METHANOCOCCI Methanococcoides Enterobacteriaceae Methanococcoides Wikipedia	1740 Kb 1444 1595 1742 1865	1	31% L77117	I-Craig Venter Institute University of Illinois at Urbana-Champaign	IBM-Annotation I-Craig Venter Institute Intard I-Craig Venter Institute NCBI	Science 272(5313):1073 1996-09-28	1996-09-28	Venter CJ
G000003	Synechocystis sp. PCC 6803	B	Cyanobacteria Taxonomy Enterobacteriaceae Cyanobacteria	3947 Kb 3442 3842 3947 4342	1	4	RA000022	Kazusa DNA Research Institute	DBI Research 3(1):9-16 1996-06-30	1996-06-30	Tabata S
G000002	Mycoplasma genitalium G37	B	Tenericutes Tenericutes Enterobacteriaceae Mycoplasma Wikipedia	580 Kb 525 560 580 605	1	32% L43957	I-Craig Venter Institute	IBM-Annotation I-Craig Venter Institute Holmgren NCBI	Science 260(5102):403 1995-10-20	1995-10-20	Fraser CM
G000001	Haemophilus influenzae Rd (Kw20)	B	PROTEOBACTERIA-GAMMA Haemophilus Enterobacteriaceae Haemophilus Wikipedia	1830 Kb 1742 1865 1870 1985	1	38% L43921 L43921	I-Craig Venter Institute	IBM-Annotation I-Craig Venter Institute Intard I-Craig Venter Institute NCBI National Lab NCBI Swiss	Science 260(5102):512 1995-07-28	1995-07-28	Venter CJ

54

(cc) BY ©2013 統合データベース講習会 Licensed Under CC 表示 2.1 E 真核生物 B 原核生物 A アーキア

## 実習4-2：メタゲノムプロジェクトを眺める

**GOLD**  
Genomes Online Database

Last update: 2013-07-08  
Total # of genomes: 27099  
Download GOLD

[Home](#)

Genome Map  
Genome Earth  
Search  
News  
Statistics  
Team  
Reference  
Contact

[Classification](#)  
• Studies: 392  
• Samples: 0

**Welcome to the Genomes OnLine Database**

GOLD: Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world.

**Metagenomes**

**Isolate Genomes**

**Genome Distribution**

• Project Type  
• Sequencing Status  
• Phylogenetic

**1. Register**  
**GOLD**  
Genomes Online Database  
Register your project information and Metadata in Genomes Online Database  
**Register**

**2. Annotate**  
**SIGS**  
Standards in Generic Sciences  
An Open Access Journal of the Generic Standards Consortium  
Annotate your microbial genome or metagenome with IMG/ER or IMG/MER  
**Annotate**

**3. Publish**  
**SIGS**  
Standards in Generic Sciences  
An Open Access Journal of the Generic Standards Consortium  
Publish your genome or metagenome in open access standards-supportive journal.  
**Publish**

U.S. DEPARTMENT OF ENERGY Office of Science

©2012 The Regents of the University of California  
Disclaimer | Credits

Version 4.0

## 実習4-2：メタゲノムプロジェクトを眺める

Metagenome Studies: 392									
Engineered: 34		Environmental: 230		Host-associated: 128					
<< first	< prev	1	2	3	4	next >	last >	(100)	first
GOLD ID	METAGENOME STUDY	SAMPLE COUNT	ECOSYSTEM	ECOSYSTEM CATEGORY	ECOSYSTEM TYPE	ECOSYSTEM SUBTYPE	SPECIFIC ECOSYSTEM	SEQUENCING CENTER	PUBLICATION
Gm00100	Human fecal microbiome from healthy Japanese men and adults, Univ of Tokyo		Host-associated	Human	Digestive system	Large intestine	Fecal	Univ of Tokyo	DNA Research 14(1):69-70 2007-12-05
Gm00113	Nasutitermes gut microbiome from Costa Rica		Host-associated	Anthropode	Digestive system	Hindgut	Unclassified	DOE Joint Genome Institute	Nature 450: 560-563 2007-11-26
Gm00112	Wastewater biofilter microbial community from Singapore and Univ of Tokyo, Japan that are treated for degrading		Engineered	Wastewater	Industrial wastewater	Unclassified	Unclassified	DOE Joint Genome Institute National University of Singapore	Lai Wen-Tso
Gm00011	Endophytic eucaryotic community of Plant roots		Host-associated	Plants	Rhizoplane	Endophytes	Unclassified	DOE Joint Genome Institute	Tringe SG
Gm00053	Marine microbial community from Chesapeake Bay		Environmental	Aquatic	Marine	Intertidal zone	Unclassified	Univ of Delaware Partners	Wommack E
Gm00073	Soil Fungal communities from forest soil		Environmental	Terrestrial	Soil	Unclassified	Unclassified	Univ of Alaska Partners	Taylor D
Gm00022	Marine viriplankton communities from surface and estuarine waters		Environmental	Aquatic	Marine	Neritic zone	Unclassified	I-Craig Venter Institute	Williamson S
Gm00069	Elephant Grass decomposing microbial community		Engineered	Solid waste	Composting	Gress	Bioreactor	DOE Joint Genome Institute	Tringe SG
Gm00068	Gut microbiome of Tupaia abdominalis		Host-associated	Anthropode	Digestive system	Unclassified	Unclassified	DOE Joint Genome Institute	Tringe SG
Gm00067	Anophelis gibratrensis gut microbiome from Wimberly, NC		Host-associated	Anthropoda	Digestive system	Unclassified	Unclassified	DOE Joint Genome Institute	Tringe SG
Gm00067	Poplar biomass bioreactor microbial communities from Brookhaven National Lab, NY		Engineered	Solid waste	Composting	Wood	Bioreactor	DOE Joint Genome Institute	Daniel van der Lelie

56

(cc) BY ©2013 統合データベース講習会 Licensed Under CC 表示 2.1

## 実習4-3：あの生物のゲノムは読まれているのか調べてみる

**GOLD**  
Genomes Online Database

Last update: 2013-07-08  
Total # of genomes: 27099  
Download GOLD: [\[PDF\]](#)

**Home**

- Home
- Genome Map
- Genome Earth
- Search**
- News
- Statistics
- Team
- Reference
- Contact

[\[Facebook\]](#) [\[Twitter\]](#) [\[Web\]](#) [\[Blogger\]](#)

**Welcome to the Genomes OnLine Database**

**GOLD**: Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world.

**Metagenomes**      **Isolate Genomes**      **Genome Distribution**

**Classification**  
 • Studies: 392  
 • Samples: 0

**Complete Projects: 6577**  
**Incomplete Projects: 20467**  
**Targeted Projects: 1687**

**Project Type**  
 • Sequencing Status  
 • Phylogenetic

**1. Register**  
**GOLD**  
 Genomes Online Database  
 Register your project information and Metadata in Genomes Online Database  
[\[Register\]](#)

**2. Annotate**  
  
 Annotate your microbial genome or metagenome with IMG/ER or IMG/MER  
[\[Annotate\]](#)

**3. Publish**  
**SIGS**  
 Standards in Genomic Sciences  
 An Open Access Journal of the Genomic Sciences Convention  
 Publish your genome or metagenome in open access standards-supportive journal.  
[\[Publish\]](#)

©2012 The Regents of the University of California  
[\[Disclaimer\]](#) [\[Credits\]](#)

U.S. DEPARTMENT OF ENERGY Office of Science 57

[\[CC BY\]](#) ©2013 統合データベース講習会 Licensed Under CC 表示 2.1

## 実習4-3：あの生物のゲノムは読まれているのか調べてみる

**SEARCH RESULTS: 37**

A Archaeal: 2 B Bacterial: 25 E Eukaryotic: 9 M Microbial: 1

hide

Google

<< first < prev 1 next > last >> 100 [\[Select\]](#)

Goldstamp	Organism	Domain	Type	Size	Contact	Project Status
GO1123	Canis latrans	E	Whole Genome Sequencing			incomplete
GO1124	Canis lupus	E	Whole Genome Sequencing			incomplete
GO0093	Clostridium colanicola 209318	B	Whole Genome Sequencing	Broad Institute		incomplete
GU1280	Clostridium colanicola DSM 13634	B	Whole Genome Sequencing	Kyrpides, Nikos		targeted

59

## 実習4-3：あの生物のゲノムは読まれているのか調べてみる

Organism Nameに学名を入れて検索

Search Advanced Search Metagenome Search Metadata Search

Select Fields You Displayed on Search Results

<input checked="" type="checkbox"/> Organism	<input checked="" type="checkbox"/> Domain	<input checked="" type="checkbox"/> Type
<input checked="" type="checkbox"/> Size	<input checked="" type="checkbox"/> Information	<input checked="" type="checkbox"/> Data-Search
<input checked="" type="checkbox"/> Sequencing Centers	<input type="checkbox"/> Funding	<input checked="" type="checkbox"/> Genome Database
<input type="checkbox"/> Publication Journal	<input type="checkbox"/> Contact	<input checked="" type="checkbox"/> Project Status
<input type="checkbox"/> GC Content	<input type="checkbox"/> Habitat	<input type="checkbox"/> Sequencing Status
<input type="checkbox"/> Sequencing Country	<input type="checkbox"/> All Fields	

Reset Default Selections Clear All Selections SEARCH GOLD

Search by:

ORGANISM INFORMATION

Domain MATCHES ALL Organism Name Canis

Phylum MATCHES ALL

Organism Name MATCHES ALL Canis

GENOME PROJECT INFORMATION

Goldstamp

Project Status MATCHES ALL

Sequencing Centers MATCHES ALL

SEQUENCING INFORMATION

Sequencing Status MATCHES ALL

Sequencing Country MATCHES ALL

submit search | reset all selections

©2011 The Regents of the University of California  
[\[Disclaimer\]](#) [\[Credits\]](#)

[\[CC BY\]](#) ©2013 統合データベース講習会 Licensed Under CC 表示 2.1

U.S. DEPARTMENT OF ENERGY Office of Science

58

## モデル生物のゲノムデータベース

[\[CC BY\]](#) ©2012 統合データベース講習会 Licensed Under CC 表示 2.1

60

## 代表的なモデル生物のゲノムデータベース

### ◆ヒト H-InvDB Annotated Human Gene Database

○ H-InvDB: H-Invitational Database  
<http://www.h-invitational.jp/>

### ◆マウス MGI

○ MGI: Mouse Genome Informatics  
<http://www.informatics.jax.org/>

### ◆ショウジョウバエ

○ FlyBase: FlyBase  
(A Database of Drosophila Genes & Genomes)  
<http://flybase.org/>



### ◆酵母 SGD Saccharomyces GENOME DATABASE

○ SGD: Saccharomyces Genome Database  
<http://www.yeastgenome.org/>

### ◆線虫

○ WormBase  
<http://www.wormbase.org>

[CC BY] ©2013 統合データベース講習会 Licensed Under CC 表示 2.1

### ◆シアノバクテリア CyanoBase

○ CyanoBase: Genome database for Cyanobacteria  
<http://genome.microbedb.jp/cyanobase>

61

## おわりに

◆DDBJ、Ensembl、BioMart、GOLD共に、時間の都合で紹介できなかった機能がたくさんありますので、統合TVなどを参考にしながらぜひ使い倒してみてください

[CC BY] ©2013 統合データベース講習会 Licensed Under CC 表示 2.1

62