

# 自己紹介

- ・ 国立遺伝学研究所 大量遺伝情報研究室 特任研究員
  - ・ NBDC 統合データベースプロジェクト
    - ・ 統合微生物DB MicrobeDB.jp開発
    - ・ CyanoBase, RhizoBase開発・運用
    - ・ 微生物ゲノムのゲノムアノテーション
    - ・ DDBJ塩基配列登録システム開発・リソースRDF化



# 質問

---

- ・ゲノムアノテーションをしたことがある人？
- ・これからゲノムアノテーションをする予定のある人？

統合データベース講習会: AJACS54

2015.06.17

# 微生物におけるNGSデータから ゲノムアノテーションまで

国立遺伝学研究所大量遺伝情報研究室

藤澤 貴智

# ゲノムプロジェクトの流れ

---

1. シーケンシング
2. アセンブル or マッピング/コンセンサス配列
3. ゲノムアノテーション
4. ゲノム情報をINSDCに登録・公開

# 微生物ゲノムアノテーションデータフローとDDBJ関連サービス

NGS sequence read



Genome sequence(s)



Genome structural  
annotation



Genome functional  
annotation



Genome annotation  
refinement



Data/Database  
publishing

## DDBJ Read Annotation Pipeline

NGSデータを入力として配列アセンブルを行なう  
自動パイプライン、Galaxyによる解析ワークフ  
ローも提供



微生物ゲノム配列に対して遺伝子予測と簡易機能ア  
ノテーションを行なう自動パイプライン

## Genome Refine

ゲノムアノテーション精度向上のための  
ウェブサービス開発

# ゲノムアノテーションの概要

---

# 目次

---

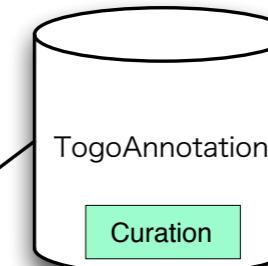
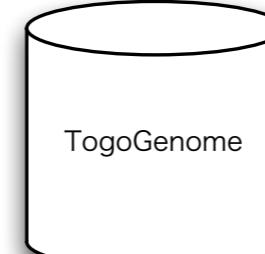
- ・ 微生物ゲノムアノテーションの概要
- ・ ゲノムアノテーション関連データベース・ツールの紹介
- ・ GenomeRefine/MicrobeDB.jpの紹介

# ゲノム（遺伝子）アノテーションの系譜

Databases for cyanobacteria and Dataflow of gene annotation, example: slr0473

INSDC

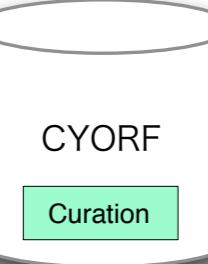
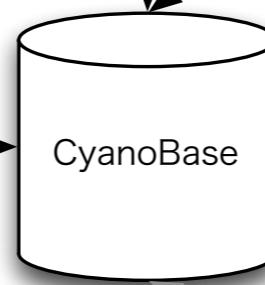
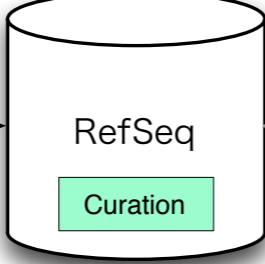
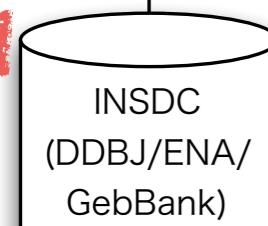
```
gene      2607812..2610058
          /gene="phy"
CDS       2607812..2610058
          /gene="phy"
          /note="ORF_ID:slr0473"
          /codon_start=1
          /transl_table=11
          /product="phytochrome"
          /protein_id="BAA10307.1"
          /db_xref="GI:1001165"
```



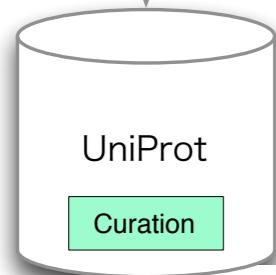
mdb:gene\_product  
 mdb:gene\_product\_literature  
 mdb:gene\_symbol  
 mdb:gene\_symbol\_literature

CyanoBase

Gene symbol	<i>cph1, hik35</i>
Gene symbol Extracted from literature	<i>cph1, hik35, phy, chk35</i>
Gene product	cyanobacterial phytochrome 1, two-component sensor histidine kinase
Gene product Extracted from literature	Cph1, SyCph1, Slr0473, Chk35, AphA, Hik35

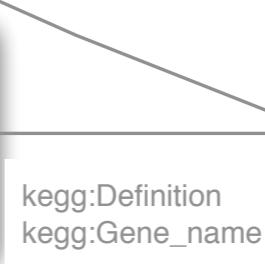
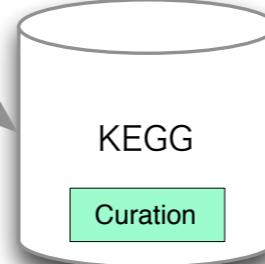


cyorf:Definition  
 cyorf:Gene\_name



RefSeq

Gene symbol	phy
Gene description	phytochrome
Locus tag	slr0473
Gene type	protein coding
RefSeq status	REVIEWED
Organism	<i>Synechocystis</i> sp. PCC 6803 (strain: PCC 6803)
Lineage	Bacteria; Cyanobacteria; Oscillatoriophycideae; Chroococcales; Synechocystis



CYORF

ORF ID: slr0473  
 Gene name: cph1, hik35  
 Definition: cyanobacterial phytochrome 1, two-component sensor histidine kinase  
 Position: 2607812..2610058  
 Length: 748 aa [AA seq] 2247 bp [NT seq] +upstream 0 bp +downstream 0 bp

CyanoBase: slr0473  
 UniProt: Q55168  
 KEGG: K11354  
 Pfam / Prosite: pf:PAS\_2, pf:GAF, pf:Phytochrome, pf:HisKA, pf:HATPase\_c, pf:HIS\_KIN, pf:PHOTOCHROME\_2

phytochrome, Phytochrome-like protein cph1 (EC 2.7.13.3) (Light-regulated histidine kinase 1), cph1; two-component system, chemotaxis family, sensor kinase Cph1 [EC:2.7.13.3]

KO: K11354  
 Pfam / Prosite: pf:PAS\_2, pf:GAF, pf:Phytochrome, pf:HisKA, pf:HATPase\_c, pf:HIS\_KIN, pf:PHOTOCHROME\_2

Ortholog: Best-best show SSDB  
 Paralog: 400 show SSDB

UniProt

KEGG

Entry	slr0473	CDS	T00004
Gene name	phy		
Definition	phytochrome		
Orthology	K11354	two-component system, chemotaxis family, sensor kinase Cph1 [EC:2.7.13.3]	
Organism	syn	<i>Synechocystis</i> sp. PCC 6803	
Pathway	syn02020	Two-component system	
Module	syn_M00510	Cph1-Rcp1 (light response) two-component regulatory system	

Protein names<sup>i</sup>: Recommended name: Phytochrome-like protein cph1 (EC:2.7.13.3)  
 Alternative name(s): Bacteriophytochrome cph1, Light-regulated histidine kinase 1

Gene names<sup>i</sup>: Name:cph1  
 Ordered Locus Names:slr0473

Organism<sup>i</sup>: *Synechocystis* sp. (strain PCC 6803 / Kazusa)

Taxonomic identifier<sup>i</sup>: 1111708 [NCBI]

Taxonomic lineage<sup>i</sup>: Bacteria > Cyanobacteria > Oscillatoriophycideae > Chroococcales > *Synechocystis* >

Proteomes<sup>i</sup>: UP000001425: Chromosome

# 国際塩基配列データベース

<http://www.ddbj.nig.ac.jp/insdc/insdc-j.html>

INSDC; International Nucleotide Sequence Database Collaboration



Data type	DDBJ Center	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive		Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequence	DDBJ	European Nucleotide Archive (ENA)	GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject



DDBJ Home Page by DDBJ is licensed under a Creative Commons 表示 2.1 日本 License

# アノテーションとは？

- アノテーション（英語：annotation、英語発音：[ænə'teis̩n]）とは、あるデータに対して関連する情報（メタデータ）を注釈として付与すること。

『フリー百科事典 ウィキペディア日本語版』より引用



# ゲノムアノテーションとは？

>chromosome

AGCTTTCACTGACTGCAACGGCAATATGTCTCTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC  
TTCTGAACCTGGTACCGCCGTGAGTAAATTAAAATTGACTTAGTCACAAACTTAACCAA  
TATAGGCATAGCGCACAGACAGATAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC  
ATTACCACCAACCATTACCAACAGGTAACGGTGCAGGCTGACCGTACAGGAAACACAGAAAAAAG  
CCCGCACCTGACAGTGCAGGCTTTTCGACCAAAGGTAAACGAGGTAACACCATTGCGAGTGTGAA  
GTTCGCGGTACATCAGTGGCAAATGCAGAACGTTCTGCGTGTGCCGATATTCTGGAAAGCAATGCC  
AGGCAGGGCAGGTGCCACCGTCCTCTGCCCGCCAAAATACCAACCACCTGGTGGCGATGATTG  
AAAAAACCATAGCGGCCAGGATGCTTACCCAAATATCAGCGATGCCAACGTATTGCCAACCTTT  
GACGGACTCGCCGCCAGCCGGTTCCCGCTGGCGCAATTGAAAACCTCGTCGATCAGGAATT  
GCCAAATAAACATGTCCTGCATGGCATTAGTTGTTGGGCAGTGCCCGGATAGCATCACGCTGCGC  
TGATTGCCGTGGCGAGAAAATGTCGATGCCATTATGCCGGCGTATTAGAACGCGCGGTACAACGT  
TACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGCATTACCTCGAATCTACCGTCGATATTGCT  
GAGTCCACCCGCCGTATTGCCAGGCCATTCCGGCTGATCACATGGTGTGATGGCAGGTTACCG  
CCGGTAATGAAAAAGGCGAACTGGTGGCTGGACGCAACGGTCCGACTACTCTGCTGCGGTGCTGGC  
TGCCTGTTACGCCGATTGTCGAGATTGGACGGACGTTGACGGGTCTATACCTGCGACCCCGCGT  
CAGGTGCCCGATGCGAGGTTGAAAGTCGATGTCCTACCAAGGAAGCGATGGAGCTTCCTACTCGGCG  
CTAAAGTTCTCACCCCCGACCAATTACCCCGATGCCAGTTCCAGATCCCTGCTGATTAAAAATAC  
CGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCAGCCCGTGTGAAGACGAATTACCGTCAAGGGC  
ATTCCAATCTGAATAACATGGCAATGTCAGCGTTCTGGTCCGGGATGAAAGGGATGGTCGGCATGG  
CGCGCGCGTCTTGCAGCGATGTCACGCCCGTATTCCGTGGTGCTGATTACGCAATCATCTCCGA  
ATACAGCATCAGTTCTGCGTCCACAAAGCGACTGTGTGCGAGCTGAACGGCAATGCAGGAAGAGTC  
TACCTGGAAGGCTTACTGGAGCCGCTGGCAGTGACGGAACGGCTGGCCATTATCTCGGTGG  
TAGGTGATGGTATGCGCACCTGCGTGGGATCTGGCGAAATTCTTGCCGACTGGCCCGCCAATAT  
CAACATTGTCGCCATTGCTCAGGGATCTGAAAC<sup>11</sup>GCTCAATCTGTCGTGGTAAATAACGATGCG

# ゲノムアノテーションとは？

>chromosome

AGCTTTCACTGACTGCAACGGCAATATGTCTCTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC  
TTCTGAACGGTTACCTGCCGTGAGTAAATTAAAATTGACTTAGGTCACTAAATACTTAACCAA  
TATAGGCATAGCGCACAGACAGATAAAATTACAGAGTACACAACATCC **ATGAAACGCATTAGCACCACC**  
**ATTACCACCAACCATTACCAACAGGTAACGGTGCAGGCTGACCGTACAGGAAACACAGAAAAAAG**  
CCCGCACCTGACAGTGCAGGCTTTTCGACCAAAGGTAAACGAGGTAACAACC **ATGCGAGTGTGAA**  
**GTTCGCGGTACATCAGTGGCAAATGCAGAACGTTCTGCGTGTGCCGATATTCTGAAAGCAATGCC**  
AGGCAGGGCAGGTGGCCACCGTCCCTCTGCCCGCCAAAATACCAACCACCTGGTGGCGATGATTG  
AAAAAAACCATTAGCGGCCAGGATGCTTACCCAATATCAGCGATGCCAACGTATTGCCCCAACTTT  
GACGGGACTCGCCGCCAGCCGGGTTCCCGCTGGCGCAATTGAAAACCTTCGTCATCAGGAATT  
GCCAAATAAAACATGCTGCATGGCATTAGTTGTTGGGCAGTGCCCGGATAGCATCACGCTGCGC  
TGATTGCGTGGCGAGAAAATGTCGATGCCATTATGCCGGCGTATTAGAACGCGCGGTACAACGT  
TACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGCATTACCTCGAATCTACCGTCGATATTGCT  
GAGTCCACCCGGTATTGGCAAGCCGCATTCCGGCTGATCACATGGGCTGATGGCAGGTTACCG  
CCGGTAATGAAAAAGGCGAACTGGTGGTGCTGGACGCAACGGTCCGACTACTCTGCTGCGGTGCTGGC  
TGCCTGTTACGCGCCGATTGTTGCGAGATTGGACGGACGTTGACGGGTCTATACCTGCGACCCCGT  
CAGGTGCCCGATGCGAGGTTGAAAGTCGATGCTACCAAGGAAGCGATGGAGCTTCTACTTCGGCG  
CTAAAGTTCTCACCCCGCACCATTACCCCGATGCCAGTTCCAGATCCCTGCTGATTAAAAATAC  
CGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCAGCCCGTGAAGACGAATTACCGTCAAGGGC  
ATTCCAATCTGAATAACATGGCAATGTTCAGCGTTCTGGTCCGGGATGAAAGGGATGGTCGGCATGG  
CGGCGCGCGTCTTGCAGCGATGTCACGCGCCGTATTCCGTGGTGCTGATTACGCAATCATCTCCGA  
ATACAGCATCAGTTCTGCGTCCACAAAGCGACTGTGTGCGAGCTGAACGGCAATGCAGGAAGAGTC  
TACCTGGAAGTGGCTTACTGGAGCCGCTGGCAGTGACGGAACGGCTGGCCATTATCTCGGTGG  
TAGGTGATGGTATGCGCACCTGCGTGGGATCTGGCGAAATTCTTGCCTGACTGGCCCGCCAATAT  
CAACATTGTCGCCATTGCTCAGGGATCTTCTGAACGCTCAATCTGTCGTGGTAAATAACGATGCG<sup>12</sup>

# ゲノムアノテーションとは？

```
>CDS 190..255
  /gene="thrL"
  /locus_tag="b0001"
  /gene_synonym="ECK0001; JW4367"
  /function="leader; Amino acid biosynthesis: Threonine"
  /GO_process="GO:0009088 - threonine biosynthetic process"
  /codon_start=1
  /transl_table=11
  /product="thr operon leader peptide"
```

GTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTCTGCGTGTGCCGATATTCTGGAAAGCAATGCC  
AGGCAGGGGCAGGTGGCCACCGTCCCTCTGCCCGCCAAAATACCCAACCACCTGGTGGCGATGATTG  
AAAAAAACCATTAGCGGCCAGGATGCTTACCCAATATCAGCGATGCCAACGTATTTGCCAACCTT  
GACGGGACTCGCCGCCAGCCGGGTTCCCGCTGGCGCAATTGAAAACCTTCGTCATCAGGAATT  
GCCCAAATAAACATGTCCTGCATGGCATTAGTTGGGGCAGTGCCGGATAGCATCACGCTGCGC  
TGATTTGCCGTGGCGAGAAAATGTCGATGCCATTATGCCGGGTATTAGAACGCGCGGTACAACGT

```
CDS 337..2799
  /gene="thrA"
  /locus_tag="b0002"
  /gene_synonym="ECK0002; Hs; JW0001; thrA1; thrA2; thrD"
  /EC_number="1.1.1.3"
  /EC_number="2.7.2.4"
  /function="enzyme; Amino acid biosynthesis: Threonine"
  /experiment="N-terminus verified by Edman degradation:
  PMID 354697,4562989"
  /GO_component="GO:0005737 - cytoplasm"
  /GO_process="GO:0009090 - homoserine biosynthetic process;
  GO:0009086 - methionine biosynthetic process; GO:0009088 -
  threonine biosynthetic process"
  /note="bifunctional: aspartokinase I (N-terminal);
  homoserine dehydrogenase I (C-terminal)"
  /codon_start=1
  /transl_table=11
  /product="Bifunctional aspartokinase/homoserine
  dehydrogenase 1"
```

TAAAAAAAAGAGTGTCTGATAGCAGC  
TTAGGTCACTAAATACTTTAACCAA  
ATCC**ATGAAACGCATTAGCACCACC**  
CGCGTACAGGAAACACAGAAAAAAAG  
AGGTAACAACC**ATGCGAGTGTGAA**

CCTCGAATCTACCGTCGATATTGCT  
ATGGTGCTGATGGCAGGTTCACCG  
CCGACTACTCTGCTGCGGTGCTGGC  
CGGGGTCTATACTGCGACCCCGCGT  
GCGATGGAGCTTCCTACTTCGGCG  
AGATCCCTTGCGATTAAAAATAC  
TGAAGACGAATTACCGGTCAAGGGC  
GGGATGAAAGGGATGGTCGGCATGG  
TGCTGATTACGCAATCATCTTCCGA  
TGAACGGCAATGCAGGAAGAGTTC  
GAACGGCTGGCCATTATCTCGGTGG  
TTGCCGCACTGGCCCCGCCAATAT

# INSDC配列エントリ形式

<http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>

The screenshot shows a web browser window with the title "GenBank Sample Record". The URL in the address bar is "www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html". The page content is titled "Sample GenBank Record". Below the title, there is a navigation bar with links to PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The main content area is titled "GenBank Flat File Format" and contains the following text:

LOCUS SCU49845 5028 bp DNA PLN 21-JUN-1999  
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.  
ACCESSION U49845  
VERSION U49845.1 GI:1293613  
KEYWORDS .  
SOURCE Saccharomyces cerevisiae (baker's yeast)  
ORGANISM Saccharomyces cerevisiae  
Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;  
Saccharomycetales; Saccharomycetaceae; Saccharomyces.  
REFERENCE 1 (bases 1 to 5028)  
AUTHORS Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.  
TITLE Cloning and sequence of REV7, a gene whose function is required for  
DNA damage-induced mutagenesis in *Saccharomyces cerevisiae*  
JOURNAL Yeast 10 (11), 1503-1509 (1994)  
PUBMED 7871890  
REFERENCE 2 (bases 1 to 5028)  
AUTHORS Roemer,T., Madden,K., Chang,J. and Snyder,M.  
TITLE Selection of axial growth sites in yeast requires Axl2p, a novel  
plasma membrane glycoprotein  
JOURNAL Genes Dev. 10 (7), 777-793 (1996)  
PUBMED 8846915  
REFERENCE 3 (bases 1 to 5028)  
AUTHORS Roemer,T.  
TITLE Direct Submission  
JOURNAL Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New  
Haven, CT, USA  
FEATURES source Location/Qualifiers  
source 1..5028  
/organism="Saccharomyces cerevisiae"  
/db\_xref="taxon:4932"  
/chromosome="IX"  
/map="9"

# GFF3形式

<http://www.sequenceontology.org/gff3.shtml>

The screenshot shows a web browser window for 'The Sequence Ontology - Resources - GFF3'. The page features a large blue header with the 'SO' logo and 'The Sequence Ontology Project' text. Below the header is a green navigation bar with links: Home, Browser, Wiki, GFF3, GVF, Resources, Software, About, Request A Term, and Site Map. The main content area has a red header 'Generic Feature Format Version 3 (GFF3)'. On the left, there's a 'Summary' section with author information (Lincoln Stein, 26 February 2013, version 1.21) and a note about the persistence of flat file formats. On the right, there's a 'News' section with a recent entry about GVF being used in clinical precision medicine. The bottom half of the page displays a sample GFF3 file content.

Home > Resources > GFF3

## Generic Feature Format Version 3 (GFF3)

### Summary

Author: Lincoln Stein  
Date: 26 February 2013  
Version: 1.21

Although there are many richer ways of representing genomic features via XML and in relational database schemas, the stubborn persistence of a variety of ad-hoc tab-delimited flat file formats defies the

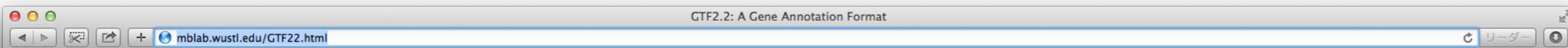
### News

► **October 2013** GVF was used in the clinical annotation of a whole genome, for precision medicine. [Integrating precision medicine in the study and clinical](#)

```
0 ##gff-version 3
1 ##sequence-region ctg123 1 1497228
2 ctg123 . gene          1000 9000 . + . ID=gene00001;Name=EDEN
3 ctg123 . TF_binding_site 1000 1012 . + . ID=tfbs00001;Parent=gene00001
4 ctg123 . mRNA          1050 9000 . + . ID=mRNA00001;Parent=gene00001;Name=EDEN
5 ctg123 . mRNA          1050 9000 . + . ID=mRNA00002;Parent=gene00001;Name=EDEN
6 ctg123 . mRNA          1300 9000 . + . ID=mRNA00003;Parent=gene00001;Name=EDEN
7 ctg123 . exon          1300 1500 . + . ID=exon00001;Parent=mRNA00003
8 ctg123 . exon          1050 1500 . + . ID=exon00002;Parent=mRNA00001;mRNA00003
```

# GTF2.2形式

<http://mblab.wustl.edu/GTF22.html>



## GTF2.2: A Gene Annotation Format

(Revised Ensembl GTF)

### Contents

- [Introduction](#)
- [GTF Field Definitions](#)
- [Examples](#)
- [Scripts and Resources](#)

### Introduction

GTF stands for Gene transfer format. It borrows from [GFF](#), but has additional structure that warrants a separate definition and format name.

Structure is as [GFF](#), so the fields are:

<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes] [comments]

Here is a simple example with 3 translated exons. Order of rows is not important.

```
381 Twinscan CDS      380    401    .    +    0    gene_id "001"; transcript_id "001.1";
381 Twinscan CDS      501    650    .    +    2    gene_id "001"; transcript_id "001.1";
381 Twinscan CDS      700    707    .    +    2    gene_id "001"; transcript_id "001.1";
381 Twinscan start_codon 380    382    .    +    0    gene_id "001"; transcript_id "001.1";
381 Twinscan stop_codon 708    710    .    +    0    gene_id "001"; transcript_id "001.1";
```

The whitespace in this example is provided only for readability. In GTF, fields must be separated by a single TAB and no white space.

[Top](#)

[GTF Field Definitions](#)

# ゲノムアノテーションとは？

- ・ ゲノム構造アノテーション (Structured annotation)
  - ・ 遺伝子の構造やリピート配列構造などを注釈
- ・ ゲノム機能アノテーション (Functional annotation)
  - ・ 遺伝子の機能を注釈
  - ・ メタデータのアノテーション
  - ・ プロジェクトやサンプル情報などのゲノム関連リソースを注釈

# ゲノムアノテーションとは？

FASTA形式



INSDC配列エントリ形式  
(DDBJ/EMBL/GenBank)

ゲノムアノテーションの公開形式で  
あるINSDC配列エントリ  
について理解を深めていきましょう

# ゲノムアノテーション関連データベース・ツールの紹介

---

# 国際塩基配列データベース

<http://www.ddbj.nig.ac.jp/insdc/insdc-j.html>

INSDC; International Nucleotide Sequence Database Collaboration



Data type	DDBJ Center	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive		Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequence	DDBJ	European Nucleotide Archive (ENA)	GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject



DDBJ Home Page by DDBJ is licensed under a Creative Commons 表示 2.1 日本 License

# INSDC配列エントリ (GenBank形式)

<http://www.ncbi.nlm.nih.gov/nuccore/CP011474.1>

LOCUS CP011474 2337451 bp DNA circular BCT 02-JUN-2015

DEFINITION *Corynebacterium pseudotuberculosis* strain 12C, complete genome.

ACCESSION CP011474

VERSION CP011474.1 GI:827029750

DBLINK BioProject: [PRJNA282769](#)

BioSample: [SAMN03577816](#)

KEYWORDS .

SOURCE *Corynebacterium pseudotuberculosis*

ORGANISM [\*Corynebacterium pseudotuberculosis\*](#)

Bacteria; Actinobacteria; Corynebacteriales; Corynebacteriaceae;  
*Corynebacterium*.

REFERENCE 1 (bases 1 to 2337451)

AUTHORS Sousa,T.J., Mariano,D.C.B., Parise,D., Parise,M.T.D., Viana,M.V.C.,  
Guimaraes,L.C., Benevides,L.J., Rocha,F.S., Bagano,P., Almeida,S.  
and Azevedo,V.

TITLE Direct Submission

JOURNAL Submitted (15-MAY-2015) General Biology, Federal University of  
Minas Gerais, Av. Antonio Carlos 6627, Pampulha, Belo Horizonte,  
Minas Gerais 31270-901, Brazil

COMMENT Source DNA and Bacteria are available from Laboratory of Cellular  
and Molecular Genetics, Federal University of Minas Gerais.

##Genome-Assembly-Data-START##

Assembly Method :: Mira v. 3.9.18

Reference-guided Assembly :: CP002924.1

Genome Coverage :: 48x

Sequencing Technology :: Ion Personal Genome Machine (PGMTM)  
System

##Genome-Assembly-Data-END##

FEATURES Location/Qualifiers

source 1..2337451  
/organism="*Corynebacterium pseudotuberculosis*"  
/mol\_type="genomic DNA"  
/strain="12C"  
/isolation\_source="abscess of sheep with CLA"  
/host="sheep"  
/db\_xref="taxon:[1719](#)"  
/country="Brazil: Petrolina,PE"  
/lat\_lon="[9.39416 S 40.5096 W](#)"  
/collection\_date="2009"

gene 1..1812  
/gene="dnaA"  
/locus\_tag="Cp12C\_0001"

CDS 1..1812  
/gene="dnaA"  
/locus\_tag="Cp12C\_0001"  
/protein\_id="1..1812"/>

# INSDC配列エントリ (GenBank形式)

```
LOCUS      CP011474          2337451 bp    DNA     circular BCT 02-JUN-2015
DEFINITION Corynebacterium pseudotuberculosis strain 12C, complete genome.
ACCESSION  CP011474
VERSION    CP011474.1  GI:827029750
DBLINK     BioProject: PRJNA282769
            BioSample: SAMN03577816
KEYWORDS   .
SOURCE     Corynebacterium pseudotuberculosis
ORGANISM   Corynebacterium pseudotuberculosis
            Bacteria; Actinobacteria; Corynebacteriales; Corynebacteriaceae;
            Corynebacterium.
REFERENCE  1 (bases 1 to 2337451)
AUTHORS   Sousa,T.J., Mariano,D.C.B., Parise,D., Parise,M.T.D., Viana,M.V.C.,
           Guimaraes,L.C., Benevides,L.J., Rocha,F.S., Bagano,P., Almeida,S.
           and Azevedo,V.
TITLE      Direct Submission
JOURNAL   Submitted (15-MAY-2015) General Biology, Federal University of
           Minas Gerais, Av. Antonio Carlos 6627, Pampulha, Belo Horizonte,
           Minas Gerais 31270-901, Brazil
COMMENT   Source DNA and Bacteria are available from Laboratory of Cellular
           and Molecular Genetics, Federal University of Minas Gerais.
```

```
##Genome-Assembly-Data-START##
Assembly Method      :: Mira v. 3.9.18
Reference-guided Assembly :: CP002924.1
Genome Coverage       :: 48x
Sequencing Technology :: Ion Personal Genome Machine (PGMTM)
                        System
```

```
##Genome-Assembly-Data-END##
FEATURES
source      Location/Qualifiers
            1..2337451
            /organism="Corynebacterium pseudotuberculosis"
            /mol_type="genomic DNA"
            /strain="12C"
            /isolation_source="abscess of sheep with CLA"
            /host="sheep"
            /db_xref="taxon:1719"
            /country="Brazil: Petrolina,PE"
            /lat_lon="9.39416 S 40.5096 W"
            /collection_date="2009"
```

```
gene        1..1812
            /gene="dnaA"
            /locus_tag="Cp12C_0001"
CDS         1..1812
            /gene="dnaA"
            /locus_tag="Cp12C_0001"
            /codon_start=1
            /transl_table=11
```

metadata

biological features

# INSDC配列エントリ (リンク情報)

<http://www.ncbi.nlm.nih.gov/nuccore/CP011474.1>

LOCUS CP011474 2337451 bp DNA circular BCT 02-JUN-2015  
DEFINITION Corynebacterium pseudotuberculosis strain 12C, complete genome.  
ACCESSION CP011474  
VERSION CP011474.1 GI:827029750  
DBLINK BioProject: [PRJNA282769](#)  
BioSample: [SAMN03577816](#)  
KEYWORDS .  
SOURCE Corynebacterium pseudotuberculosis  
ORGANISM [Corynebacterium pseudotuberculosis](#)  
Bacteria; Actinobacteria; Corynebacteriales; Corynebacteriaceae;  
Corynebacterium.  
REFERENCE 1 (bases 1 to 2337451)  
AUTHORS Sousa,T.J., Mariano,D.C.B., Parise,D., Parise,M.T.D., Viana,M.V.C.,  
Guimaraes,L.C., Benevides,L.J., Rocha,F.S., Bagano,P., Almeida,S.  
and Azevedo,V.  
TITLE Direct Submission  
JOURNAL Submitted (15-MAY-2015) General Biology, Federal University of  
Minas Gerais, Av. Antonio Carlos 6627, Pampulha, Belo Horizonte,  
Minas Gerais 31270-901, Brazil  
COMMENT Source DNA and Bacteria are available from Laboratory of Cellular  
and Molecular Genetics, Federal University of Minas Gerais.  
  
##Genome-Assembly-Data-START##  
Assembly Method :: Mira v. 3.9.18  
Reference-guided Assembly :: CP002924.1  
Genome Coverage :: 48x  
Sequencing Technology :: Ion Personal Genome Machine (PGMTM)  
System  
##Genome-Assembly-Data-END##  
FEATURES Location/Qualifiers  
source 1..2337451  
/organism="Corynebacterium pseudotuberculosis"  
/mol\_type="genomic DNA"  
/strain="12C"  
/isolation\_source="abscess of sheep with CLA"  
/host="sheep"  
/db\_xref="taxon:[1719](#)"  
/country="Brazil: Petrolina,PE"  
/lat\_lon="[9.39416 S 40.5096 W](#)"  
/collection\_date="2009"  
gene 1..1812  
/gene="dnaA"  
/locus\_tag="Cp12C\_0001"  
1..1812  
/gene="dnaA"  
/locus\_tag="Cp12C\_0001"  
/codon\_start=1

BioProject

BioSample

リンクされたDBエントリ  
も登録 or 既存のIDのアサ  
インを行ないます

Taxonomy

locus\_tag prefix

\*BioProjectにおいて登録

# 【実習】データベースの確認

- [www.ncbi.nlm.nih.gov/nuccore/CP011474.1](http://www.ncbi.nlm.nih.gov/nuccore/CP011474.1)
  - から BioSample, BioProject、Taxonomyのエントリーへ飛び各データベースを俯瞰してみましょう。

# BioProject

<http://www.ncbi.nlm.nih.gov/bioproject/PRJNA282769>

Corynebacterium pseudotuberculosis strain:12C (ID 282769) – BioProject – NCBI

NCBI Resources How To

BioProject BioProject Advanced Search Help

**Display Settings:** Send to:

**Corynebacterium pseudotuberculosis strain:12C** Accession: PRJNA282769 ID: 282769

**Corynebacterium pseudotuberculosis genome sequencing 12C**

The goal of this study is comparative analysis of genome with other Corynebacterium pseudotuberculosis strain.

**Project Type:** Genome sequencing; **Locus Tag Prefix:** Cp12C

**Attributes:** Scope: Monoisolate; Material: Genome; Capture: Whole; Method type: Sequencing

**Relevance:** Agricultural

**Project Data:**

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (Genomic DNA)	1
Protein Sequences	2119
OTHER DATASETS	
BioSample	1
Assembly	1

**Assembly details:**

Assembly	Level	Chrs	BioSample	Taxonomy
GCA_001017615.1	Complete genome	1	SAMN03577816	Corynebacterium pseudotuberculosis

**Lineage:** Bacteria; Actinobacteria; Corynebacteriales; Corynebacteriaceae; Corynebacterium; Corynebacterium pseudotuberculosis [Taxonomy ID: 1719]

**Submission:**  
Registration date: 2-Jun-2015  
Federal University of Minas Gerais

**Related information**

See Genome Information for Corynebacterium pseudotuberculosis

NAVIGATE ACROSS  
41 additional projects are related by organism.

**Recent activity**

Turn Off Clear

- Corynebacterium pseudotuberculosis strain:12C BioProject
- Corynebacterium pseudotuberculosis strain 12C, complete genome Nucleotide
- Prokaryotic Genome Annotation Pipeline - The NCBI Handbook
- locus\_tag prefix (5) Books
- locus\_tag (2) BioProject

See more...

# Locus\_tag

[http://www.ddbj.nig.ac.jp/sub/locus\\_tag-j.html](http://www.ddbj.nig.ac.jp/sub/locus_tag-j.html)

The screenshot shows a web browser window with the following details:

- Title Bar:** /locus\_tag qualifier の記載法 | DDBJ
- Address Bar:** www.ddbj.nig.ac.jp/sub/locus\_tag-j.html
- Content Area:**
  - Section Header:** /locus\_tag qualifier の記載法
  - Section:** 背景 (Background)
    - /locus\_tag qualifier は、2003年に導入されました。その導入当初は、ゲノムプロジェクトがその配列データを更新する際などに feature 繙承をするための追跡用 ID として自由度の高い記載を可能としていました。

しかし、The American Society for Microbiology からの要請を受けて、2005年の国際実務者会議において、/locus\_tag qualifier の用法が再検討されました。その結果、/locus\_tag qualifier を恒久的に一意な ID として維持していくことを目指して、配列データの登録時に当該ゲノム専用の prefix を割り当てることにより、/locus\_tag を記載するように規則が変更されました。

DDBJにおいては、ゲノム配列データ登録用に 大量登録システム (MSS) をご用意しております。MSS申し込みフォームにおいて、適宜、指定していただければ、/locus\_tag prefix 割り当てを検討いたします。
    - Section Header:** ゲノム配列データの登録における /locus\_tag の適切な用法
      - 国際実務者会議 (International Collaborators Meeting)において、ゲノムプロジェクトを INSDC に登録するように求めていく、と合意しています。各ゲノムプロジェクトに ID を割り当てることにより、複数の配列データを各ゲノムプロジェクトに関連付けることが可能になります。この Project ID は、DDBJ フラットファイルにおいては ACCESSION 行と VERSION 行の下に表示されます。ゲノムプロジェクトの登録は DDBJ, EMBL-Bank/EBI, GenBank/NCBI で行うことができます。登録者はゲノムプロジェクトの登録に際し、同時に /locus\_tag prefix の登録を行うことができます。

/locus\_tag はゲノム上の全ての gene に体系的に割り当てる識別子(identifier, ID)であり、生物学関連団体による遺伝子名に代わる ID になります。2組の異なるゲノムの登録者が全く異なる2つのゲノムにおいて全く異なる2つの遺伝子に同じ体系に扱い名称を用いたならば、混乱を招くことになるでしょう。このようなことが起こることを防ぐために INSD (DDBJ/EMBL/GenBank) では /locus\_tag prefix を登録する仕組みを作りました。真核生物でも原核生物でもゲノムの登録者は、そのゲノム登録に先立って prefix を登録してください。そして、複数の染色体、プラスミドといったプロジェクトの全ての構成要素に同じ /locus\_tag prefix を使用してください。

/locus\_tag の prefix には英数字のみを使用し、少なくとも3文字以上でなければなりません。最初の1文字目は英字で始めますが、2文字目以降は数字でも構いません (例: A1C)。prefixには "-" " " "\*" といったシンボル記号は使用しないでください。/locus\_tag においては prefix と tag の値はアンダースコア "\_" によって区切れます (例: A1C\_00001)。

/locus\_tags は、全てのタンパク質コード遺伝子とタンパク質をコードしない RNA 遺伝子に割り当ててください。/locus\_tag はゲノム配列の登録において mRNA, CDS, 5'UTR, 3'UTR, intron, exon, tRNA, rRNA, ncRNA, misc\_RNA, などの feature に記載します。repeat\_region には /locus\_tag qualifier を記載しないでください。同じ値を持つ /locus\_tag はある単一の gene の全ての構成要素に使用します。例えば、ある特定の gene を示す exon, CDS, mRNA といった全ての feature には同じ値を持つ /locus\_tag を記載します。また、1つの /locus\_tag には1つの /gene qualifier が対応するようにしてください。すなわち、もし何れかの feature においてある /locus\_tag がある /gene qualifier と対応している場合は、その /gene qualifier で示される遺伝子シンボルのみが、その /locus\_tag を含む他の全ての feature にも存在していなければなりません。

/locus\_tag はゲノム内の gene に体系的に記載してください。一般的には、ゲノムでの出現順序になることが期待されます。登録者がゲノム配列とその annotation を更新した場合、新規の gene は、[用例 1] その次に続く使用可能な locus\_tag、または、[用例 2] 登録者は最初の locus\_tag 割り当ての際に予め gap を残しておくことも可能なので new\_annotation の際にこの gap を埋めるような値を記載すること、の何れかが可能です。

# BioSample

<http://www.ncbi.nlm.nih.gov/biosample/SAMN03577816>

MIGS Cultured Bacterial/Archaeal sample from Corynebacterium pseudotuberculosis 12C – BioSample – NCBI

NCBI Resources How To fujisawa My NCBI Sign Out

BioSample BioSample Advanced Search Help

Display Settings: Full Send to:

**MIGS Cultured Bacterial/Archaeal sample from Corynebacterium pseudotuberculosis 12C**

Identifiers BioSample: SAMN03577816; Sample name: Complete genome of *Corynebacterium pseudotuberculosis* 12C

Organism [Corynebacterium pseudotuberculosis](#)  
cellular organisms; Bacteria; Actinobacteria; Actinobacteria; Corynebacteriales; Corynebacteriaceae; Corynebacterium

Package [MIGS: cultured bacteria/archaea, host-associated; version 4.0](#)

Attributes

<b>strain</b>	12C
<b>collection date</b>	2009
<b>environment biome</b>	-
<b>environment feature</b>	-
<b>environment material</b>	abscess of sheep with CLA
<b>geographic location</b>	<a href="#">Brazil: Petrolina, PE</a>
<b>host</b>	sheep
<b>isolation and growth condition</b>	abscess of sheep with CLA
<b>latitude and longitude</b>	<a href="#">9.39416 S 40.5096 W</a>
<b>number of replicons</b>	1
<b>reference for biomaterial</b>	Genome complete of <i>Corynebacterium pseudotuberculosis</i> Strain 12C
<b>observed biotic relationship</b>	parasite
<b>estimated size</b>	2,34
<b>host disease</b>	cutaneous lesion
<b>source material identifiers</b>	Biological sample

Description Keywords: GSC:MIxS;MIGS:4.0

BioProjects [PRJNA282769](#) *Corynebacterium pseudotuberculosis* strain:12C  
Retrieve [all samples](#) from this project

[PRJNA224116](#)  
Retrieve [all samples](#) from this project

Submission [Federal University of Minas Gerais, Thiago Sousa; 2015-04-30](#)

Related information

BioProject

Nucleotide

Assembly

Taxonomy

Recent activity

Turn Off Clear

MIGS Cultured Bacterial/Archaeal sample from *Corynebacterium pseudotuberculosis* biosample

*Corynebacterium pseudotuberculosis* strain 12C, complete genome Nucleotide

*Corynebacterium pseudotuberculosis* strain:12C BioProject

Prokaryotic Genome Annotation Pipeline - The NCBI Handbook

locus\_tag prefix (5) Books

See more...

# BioSample/Attributes

<http://trace.ddbj.nig.ac.jp/biosample/attribute.html?all=all>

BioSample Sample Attribute

trace.ddbj.nig.ac.jp/biosample/attribute.html?all=all

DDBJ  
DNA Data Bank of Japan

BioSample

Login & Submit | Databases | English | Contact

Google™カスタム検索

Home | Handbook | Sample Attribute | FAQ | Search | About BioSample

HOME > Sample Attribute

サンプル属性

List all sample attributes

Sample type (Core Package)

Genome, metagenome or marker sequences (MixS compliant)  
 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。

Name	Description (Japanese)	Description	Value format
sample_name	sample name は登録者がサンプルに付ける名前で ID として機能します。sample name は Submission において <b>ユニーク</b> である必要があります。<登録後は sample name を変更することはできません。1 submission 中で投稿された複数サンプルは sample_name のアルファベット順に並べ替えられますのでご注意ください。	The sample name is a name that you choose for the sample, it works as an ID. Each sample name must be <b>unique</b> in a submission. <b>Sample name cannot be changed after registration.</b> After submitting BioSamples from D-way, BioSamples are sorted in alphabetical order of sample_name values	[A-Z][a-z][0-9] 000+_-
sample_title	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 <b>ユニーク</b> 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.	Sample title should be short and informative. Each sample title must be <b>unique</b> in a submission. Examples: 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.	
description	サンプルに対する簡潔な補足情報。	A brief description for the sample.	
organism	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。	The most descriptive organism name for this sample (to the species, if relevant) in the NCBI Taxonomy database. If it is not in the database, provide as much information about the organism as possible and the DDBJ staff apply a new TaxID to NCBI Taxonomy.	
taxonomy_id	NCBI Taxonomy identifier. 個別の生物、 <b>メタゲノム</b> 、環境サンプルに割り当てられています。	NCBI Taxonomy identifier. This is appropriate for individual organisms, <b>some metagenomes and environmental samples</b> .	{integer}
bioproject_id	関連する BioProject アクセション番号 (PRJDB)	Associated BioProject accession number (PRJDB)	PRJDB[0-9]+
locus_tag_prefix	アノテーションを記載するゲノムに対するユニークな locus_tag_prefix /locus tag について	A locus tag prefix for an annotated genome /locus tag qualifier	
strain	微生物や真核生物の株名。	Microbial/eukaryotic strain name	
breed	品種の名称。主に家畜化された動物に用いられます。	Breed name - chiefly used in domesticated animals or plants	
cultivar	植物の栽培品種名。	Cultivar name - cultivated variety of plant	
isolate	サンプルの得られた individual isolate	Identification or description of the specific individual from which this sample was obtained	
label	単離されたサンプルのラベル名、もしくは個々の動物の名前 (例: Clint)。	A label for sample, or name of an individual animal (e.g., Clint)	
biomaterial_provider	実験材料の提供元(例: culture collection identifier, Principal Investigator, 研究室名)	name and address of the lab or PI, or a culture collection identifier	
collection_date	サンプルを採取した日付 (例: 2008-01-23T19:23:10, 2008-01-23, 2008-01-20, 1952-10-21T11:43Z/1952-10-	Time of sampling (single instance or interval, eg., 2008-01-23T19:23:10, 2008-01-23, 2008-01-20, 1952-10-	{timestamp}

# 国際塩基配列データベース

INSDC; International Nucleotide Sequence Database Collaboration



Data type	DDBJ Center	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive		Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequence	DDBJ	European Nucleotide Archive (ENA)	GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject



# Taxonomy Database

<http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1719>

Taxonomy browser (Corynebacterium pseudotuberculosis)

NCBI Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search for  as   lock

Display 3 levels using filter: none

Nucleotide  Nucleotide EST  Nucleotide GSS  Protein  Structure  Genome  Popset  SNP  
 Domains  GEO Datasets  UniGene  PubMed Central  Gene  HomoloGene  SRA Experiments  MapView  
 LinkOut  BLAST  TRACE  Probe  Assembly  Bio Project  Bio Sample  Bio Systems  
 Clone DB  dbVar  Epigenomics  GEO Profiles  PubChem BioAssay  Protein Clusters  Host

Lineage (full): root; cellular organisms; Bacteria; Actinobacteria; Actinobacteria; Corynebacteriales; Corynebacteriaceae; Corynebacterium

o [Corynebacterium pseudotuberculosis](#) Click on organism name to get more information.

- [Corynebacterium pseudotuberculosis 1/06-A](#)
- [Corynebacterium pseudotuberculosis 1/06-B](#)
- [Corynebacterium pseudotuberculosis 1/06-C](#)
- [Corynebacterium pseudotuberculosis 1/06-D](#)
- [Corynebacterium pseudotuberculosis 1/06-E](#)
- [Corynebacterium pseudotuberculosis 1/06-F](#)
- [Corynebacterium pseudotuberculosis 1/06-G](#)
- [Corynebacterium pseudotuberculosis 1/06-H](#)
- [Corynebacterium pseudotuberculosis 1/06-I](#)
- [Corynebacterium pseudotuberculosis 1/06-J](#)
- [Corynebacterium pseudotuberculosis 1/06-K](#)
- [Corynebacterium pseudotuberculosis 1/06-L](#)
- [Corynebacterium pseudotuberculosis 1/06-M](#)
- [Corynebacterium pseudotuberculosis 1/06-N](#)
- [Corynebacterium pseudotuberculosis 1/06-O](#)
- [Corynebacterium pseudotuberculosis 1/06-P](#)
- [Corynebacterium pseudotuberculosis 1/06-Q](#)
- [Corynebacterium pseudotuberculosis 1/06-R](#)
- [Corynebacterium pseudotuberculosis 1/06-S](#)
- [Corynebacterium pseudotuberculosis 1/06-T](#)
- [Corynebacterium pseudotuberculosis 1/06-U](#)
- [Corynebacterium pseudotuberculosis 1/06-V](#)
- [Corynebacterium pseudotuberculosis 1/06-W](#)
- [Corynebacterium pseudotuberculosis 1/06-X](#)
- [Corynebacterium pseudotuberculosis 1/06-Y](#)
- [Corynebacterium pseudotuberculosis 1/06-Z](#)

Disclaimer: The NCB information.

Comments and questions

Taxonomy browser (Corynebacterium pseudotuberculosis)

NCBI Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search for  as   lock

Display 3 levels using filter: none

## Corynebacterium pseudotuberculosis

Taxonomy ID: 1719  
Inherited blast name: high GC Gram+  
Rank: species  
Genetic code: [Translation table 11 \(Bacterial, Archaeal and Plant Plastid\)](#)  
Other names:

synonym: [Mycobacterium tuberculosis-ovis](#)  
synonym: [Corynebacterium pseudotuberculosis-ovis](#)  
synonym: [Corynebacterium preisz-nocardi](#)  
synonym: [Corynebacterium ovis](#)  
synonym: [Bacillus pseudotuberculosis-ovis](#)  
synonym: [Bacillus pseudotuberculosis](#)  
authority: [Corynebacterium pseudotuberculosis \(Buchanan 1911\) Eberson 1918](#)  
authority: "Mycobacterium tuberculosis-ovis" Krasil'nikov 1941  
authority: "Corynebacterium pseudotuberculosis-ovis" (Lehmann and Neumann 1896) Hauduroy et al. 1937  
authority: "Corynebacterium preisz-nocardi" Hauduroy et al. 1937  
authority: "Corynebacterium ovis" Bergey et al. 1923  
authority: "Bacillus pseudotuberculosis-ovis" Lehmann and Neumann 1896  
authority: "Bacillus pseudotuberculosis" Buchanan 1911

type material: [NCTC 3450](#)  
type material: [NBRC 15363](#)

Entrez records		
Database name	Subtree links	Direct links
Nucleotide	<a href="#">166</a>	<a href="#">108</a>
Nucleotide GSS	<a href="#">1,226</a>	<a href="#">1,226</a>
Protein	<a href="#">58,806</a>	<a href="#">25,473</a>
Genome	<a href="#">1</a>	<a href="#">1</a>
Popset	<a href="#">5</a>	<a href="#">5</a>
PubMed Central	<a href="#">84</a>	<a href="#">84</a>
Gene	<a href="#">33,033</a>	<a href="#">1</a>
SRA Experiments	<a href="#">7</a>	<a href="#">6</a>
Assembly	<a href="#">26</a>	<a href="#">9</a>
Bio Project	<a href="#">44</a>	<a href="#">14</a>
Bio Sample	<a href="#">42</a>	<a href="#">27</a>
Bio Systems	<a href="#">2,961</a>	<a href="#">502</a>
Protein Clusters	<a href="#">2,106</a>	-
Taxonomy	<a href="#">16</a>	<a href="#">1</a>

# NCBI Taxonomy Database

---

- NCBI Taxonomyは生物種の名称とその分類体系に関して整理されたデータベース
- NCBIの生物種リソースのハブであり、さらには国際塩基配列データベース（DDBJ/EMBL/GenBank）に登録されている塩基配列が由来する生物種名とその分類は、ここで一元管理

# 【実習】Taxonomy IDを調べる

- Taxonomy Databaseで生物種名“*Synechocystis* sp. PCC 6803”からTaxonomy IDを検索する
- また、下位の階層のtaxonomy IDやrank情報を確認して下さい。

# 検索結果

Taxonomy browser (Synechocystis sp. PCC 6803)

NCBI Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search for **Synechocystis sp. PCC6803** as complete name  lock

Display 3 levels using filter: none

Nucleotide  Nucleotide EST  Nucleotide GSS  Protein  Structure  Genome  Popset  SNP  
 Domains  GEO Datasets  UniGene  PubMed Central  Gene  HomoloGene  SRA Experiments  MapView  
 LinkOut  BLAST  TRACE  Probe  Assembly  Bio Project  Bio Sample  Bio Systems  
 Clone DB  dbVar  Epigenomics  GEO Profiles  PubChem BioAssay  Protein Clusters  Host

**Lineage (full):** root; cellular organisms; Bacteria; Cyanobacteria; Oscillatoriophycideae; Chroococcales; Synechocystis

◦ **Synechocystis sp. PCC 6803** [LinkOut](#) Click on organism name to get more information.

- [Synechocystis sp. PCC 6803 substr. GT-I](#) [LinkOut](#)
- [Synechocystis sp. PCC 6803 substr. GT-S](#) [LinkOut](#)
- [Synechocystis sp. PCC 6803 substr. Kazusa](#) [LinkOut](#)
- [Synechocystis sp. PCC 6803 substr. PCC-N](#) [LinkOut](#)
- [Synechocystis sp. PCC 6803 substr. PCC-P](#) [LinkOut](#)

**Disclaimer:** The NCBI taxonomy database is not an authoritative source for nomenclature or classification - please consult the relevant scientific literature for the most reliable information.

Comments and questions to [\[Help\]](#)

Taxonomy browser (Synechocystis sp. PCC 6803)

NCBI Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy

Search for  as complete name  lock

Display 3 levels using filter: none

**Synechocystis sp. PCC 6803**

**Taxonomy ID:** 1148  
**Inherited class name:** cyanobacteria  
**Rank:** species

**Genetic code:** [Translation table 11 \(Bacterial, Archaeal and Plant Plastid\)](#)

**Other names:**

- synonym: [Synechocystis sp. ATCC 27184](#)
- synonym: [Synechocystis sp. \(ATCC 27184\)](#)
- synonym: [Aphanocapsa sp. N-1](#)
- synonym: [Aphanocapsa sp. \(strain N-1\)](#)
- includes: [Synechocystis sp. PCC 6803 B](#)
- includes: [Synechocystis sp. PCC 6803 A](#)
- equivalent name: [Synechocystis sp. \(strain PCC 6803\)](#)
- equivalent name: [Synechocystis sp. \(PCC 6803\)](#)

[Lineage \(full \)](#)

Entrez records	
Database name	Subtre
Nucleotide	
Protein	2
Structure	
Genome	
GEO Datasets	
PubMed Central	
Gene	
SRA Experiments	
Probe	
Assembly	
Bio Project	
Bio Sample	
Bio Systems	

# Taxonomy Database

The screenshot shows the NCBI Taxonomy Browser interface. At the top, there's a navigation bar with links to Entrez, PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and Books. Below the navigation bar is a search bar with the query "Synechocystis sp. PCC6803". To the right of the search bar are dropdown menus for "as" (set to "complete name") and "lock" (checked), and buttons for "Go" and "Clear". Below the search bar is a "Display" button followed by a dropdown menu set to "3 levels using filter: none". A red arrow points from the text above to this dropdown menu.

The "Token set" option returns longer names that include the search terms, e.g., hybrid taxa. See what happens if you query "Bos taurus" using the "Complete match" option versus the "Set of tokens" option. The "Phonetic search" option can be used when you are not sure about the exact spelling of a organism name. It tries to find the phonetically closest strings (try "Drozofila" as an example).

**This is the top level of the taxonomy database maintained by NCBI/GenBank. You can explore any of the taxa listed below by clicking it.**

- [Archaea](#)
- [Bacteria](#)
- [Eukaryota](#)
- [Viroids](#)
- [Viruses](#)
- [Other](#)
- [Unclassified](#)

**These are direct links to some of the organisms commonly used in molecular research projects:**

[Arabidopsis thaliana](#)

[Bos taurus](#)

[Caenorhabditis elegans](#)

[Chlamydomonas reinhardtii](#)

[Danio rerio \(zebrafish\)](#)

[Dictyostelium discoideum](#)

[Drosophila melanogaster](#)

[Escherichia coli](#)

[Hepatitis C virus](#)

[Homo sapiens](#)

[Mus musculus](#)

[Mycoplasma pneumoniae](#)

[Oryza sativa](#)

[Plasmodium falciparum](#)

[Pneumocystis carinii](#)

[Rattus norvegicus](#)

[Saccharomyces cerevisiae](#)

[Schizosaccharomyces pombe](#)

[Takifugu rubripes](#)

[Xenopus laevis](#)

[Zea mays](#)

**Disclaimer:** The NCBI taxonomy database is not an authoritative source for nomenclature or classification - please consult the relevant scientific literature for the most reliable information.

# Taxonomy ID

<http://trace.ddbj.nig.ac.jp/news/2013-12-13.html>

生物の株情報を管理する方法が変更になります

**DDBJ**  
DNA Data Bank of Japan

**News**

TOP > News

[SHARE](#)

« 前の News | 次の News »

**Follow us on**

[Twitter](#) [RSS](#) [YouTube](#)

**Archives**

News by year: [Latest](#) ▾

[FAQs](#)

[Handbooks](#)

**生物の株情報を管理する方法が変更になります**

作成日: 2013年12月13日; 最終更新日: 2014年12月15日

INSDC メンバーは strain-level taxonomy ID をゲノム配列が登録される微生物に対して発行していましたが、2014年2月にこの慣行を停止します。以降は strain-level taxonomy ID に替わり BioSample ID が微生物ゲノムを識別することになります。

詳細は NCBI News をご覧ください: "Planned change in bacterial strain-level information management"

**Taxonomy**

Strain-level TaxID の発行は2014年2月に停止されますが、既存 species レベルまで同定されていない生物のための informal strain を使った species-level informal name が名称として使用されます。

**BioSample**

DDBJ センターは2014年2月から BioSample の受付を開始します。対象の生物を BioSample に登録することが必須です。BioSample collection や isolation に関する情報を必要に応じて含めます。

**BioProject**

ゲノム配列を登録する場合、シークエンスプロジェクトを BioProject として登録します。BioProject アクセッション番号とゲノム配列が1対1に対応して登録される「数の微生物株」や「いくつかの薬剤抵抗性を示す微生物種」などを BioProject として登録することができます。

**Before February, 2014**

**After February, 2014**

Species-level taxonomy ID: 100  
Strain-level Taxonomy ID から BioSample ID へ

# INSDC配列エントリ (構造アノテーション)

LOCUS CP011474 2337451 bp DNA circular BCT 02-JUN-2015  
DEFINITION Corynebacterium pseudotuberculosis strain 12C, complete genome.  
ACCESSION CP011474  
VERSION CP011474.1 GI:827029750  
DBLINK BioProject: [PRJNA282769](#)  
BioSample: [SAMN03577816](#)  
KEYWORDS .  
SOURCE Corynebacterium pseudotuberculosis  
ORGANISM [Corynebacterium pseudotuberculosis](#)  
Bacteria; Actinobacteria; Corynebacteriales; Corynebacteriaceae;  
Corynebacterium.  
REFERENCE 1 (bases 1 to 2337451)  
AUTHORS Sousa,T.J., Mariano,D.C.B., Parise,D., Parise,M.T.D., Viana,M.V.C.,  
Guimaraes,L.C., Benevides,L.J., Rocha,F.S., Bagano,P., Almeida,S.  
and Azevedo,V.  
TITLE Direct Submission  
JOURNAL Submitted (15-MAY-2015) General Biology, Federal University of  
Minas Gerais, Av. Antonio Carlos 6627, Pampulha, Belo Horizonte,  
Minas Gerais 31270-901, Brazil  
COMMENT Source DNA and Bacteria are available from Laboratory of Cellular  
and Molecular Genetics, Federal University of Minas Gerais.

```
##Genome-Assembly-Data-START##  
Assembly Method :: Mira v. 3.9.18  
Reference-guided Assembly :: CP002924.1  
Genome Coverage :: 48x  
Sequencing Technology :: Ion Personal Genome Machine (PGMTM)  
System
```

```
##Genome-Assembly-Data-END##
```

FEATURES Location/O qualifiers

source 1..2337451  
/organism="Corynebacterium pseudotuberculosis"  
/mol\_type="genomic DNA"  
/strain="12C"  
/isolation\_source="abscess of sheep with CLA"  
/host="sheep"  
/db\_xref="taxon:[1719](#)"  
/country="Brazil: Petrolina,PE"  
/lat\_lon="[9.39416 S 40.50000 W](#)"  
/collection\_date="2009"

gene 1..1812  
/gene="dnaA"

CDS 1..1812  
/gene="dnaA"

/locus\_tag="Cp12C\_0001"

/codon\_start=1

/transl\_table=[11](#)

## Features/Locations

# 構造アノテーションの注釈

Location の記述法 | DDBJ

www.ddbj.nig.ac.jp/sub/ref9-j.html

リーダー

Google™ カスタム検索 Search

DDBJ DNA Data Bank of Japan

塩基配列の登録 プロジェクトの登録 塩基配列登録の前に Flat File の説明 お問い合わせ

HOME > データ登録 > 塩基配列の登録 > Location の記述法 最終更新日：2014.11.21.

## Location の記述法

国際塩基配列データベースでは、配列上の Feature の位置情報 (以下、Location) を以下のルールで記述しています。

Location記述例	意味
340..565	登録配列の340番目から565番目までの連続した領域
complement(261..457)	登録配列の261番目から457番目までの連続した相補鎖の領域
<345..500	登録配列には、そのFeatureの開始点が含まれない。 あるいは、345番目の塩基より5'側に関して、そのFeatureの正確な開始点が不明である。
345..>500	登録配列には、そのFeatureの終了点が含まれない。 あるいは、500番目の塩基より3'側に関して、そのFeatureの正確な終了点が不明である。
<345..>500	登録配列には、そのFeatureの開始点、終了点が共に含まれない。 あるいは、345番目の塩基より5'側と500番目の塩基より3'側に関して、そのFeatureの正確な開始点、終了点が共に不明である。
join(12..78,134..202)	登録配列の12番目から78番目、ならびに134番目から202番目までの不連続な領域
complement(join(12..78,134..202))	登録配列の12番目から78番目、ならびに134番目から202番目までの不連続な相補鎖の領域
467	登録配列の467番目の塩基
123^124	登録配列の123番目と124番目の間

以下の Location の記述法は、[塩基配列登録システム](#) では入力することができません。  
塩基配列登録システムの "Submission Information" の入力欄にその旨を説明していただくか、[Mass Submission System](#)をご利用下さい。

Location 記述例	意味
join(AB000000.1:100..202,134..222)	関連するエントリーAB000000 (version 1) の配列の100番目から202番目までの領域、ならびに当該エントリーの134番目から222番目までの不連続な領域

これらの Location 記述法を組み合わせることによって、これ以外の様々な位置情報を記述することができます。

より詳しい Location の記述ルールに関する情報は、[Feature Table Definition: 3.4 Location](#) をご参照ください。

# INSDC配列エントリ (GenBank形式)

```
LOCUS      CP011474          2337451 bp    DNA     circular BCT 02-JUN-2015
DEFINITION Corynebacterium pseudotuberculosis strain 12C, complete genome.
ACCESSION  CP011474
VERSION    CP011474.1  GI:827029750
DBLINK     BioProject: PRJNA282769
            BioSample: SAMN03577816
KEYWORDS   .
SOURCE     Corynebacterium pseudotuberculosis
ORGANISM   Corynebacterium pseudotuberculosis
            Bacteria; Actinobacteria; Corynebacteriales; Corynebacteriaceae;
            Corynebacterium.
REFERENCE  1 (bases 1 to 2337451)
AUTHORS   Sousa,T.J., Mariano,D.C.B., Parise,D., Parise,M.T.D., Viana,M.V.C.,
           Guimaraes,L.C., Benevides,L.J., Rocha,F.S., Bagano,P., Almeida,S.
           and Azevedo,V.
TITLE      Direct Submission
JOURNAL   Submitted (15-MAY-2015) General Biology, Federal University of
           Minas Gerais, Av. Antonio Carlos 6627, Pampulha, Belo Horizonte,
           Minas Gerais 31270-901, Brazil
COMMENT   Source DNA and Bacteria are available from Laboratory of Cellular
           and Molecular Genetics, Federal University of Minas Gerais.
```

```
##Genome-Assembly-Data-START##
Assembly Method      :: Mira v. 3.9.18
Reference-guided Assembly :: CP002924.1
Genome Coverage       :: 48x
Sequencing Technology :: Ion Personal Genome Machine (PGMTM)
                        System
```

```
##Genome-Assembly-Data-END##
FEATURES      Location/Qualifiers
source        1..2337451
              /organism="Corynebacterium pseudotuberculosis"
              /mol_type="genomic DNA"
              /strain="12C"
              /isolation_source="abscess of sheep with CLA"
              /host="sheep"
              /db_xref="taxon:1719"
              /country="Brazil: Petrolina,PE"
              /lat_lon="9.39416 S 40.5096 W"
              /collection_date="2009"
```

```
gene          1..1812
              /gene="dnaA"
              /locus_tag="Cp12C_0001"
```

```
CDS           1..1812
              /gene="dnaA"
              /locus_tag="Cp12C_0001"
              /codon_start=1
              /transl_table=11
```

## Features/Qualifiers

# 機能アノテーションの注釈

- 生物学的な特徴を、feature key（特徴を表す項目）および Qualifier（特徴をさらに特定する項目）を用いて注釈
- 配列の特徴を記述するための feature key は、
  - (1) 由来生物の特徴を記述するための feature key (source)
  - (2) 配列の中の一定の領域がもつ生物学的機能を記述するための feature key (e.g. CDS, rRNA, etc.)
  - (3) 配列の差違や変更を記述するための feature key (e.g. variation, conflict, etc.)

# 機能アノテーションの注釈: Feature/Qualifier

<http://www.ddbj.nig.ac.jp/sub/ref7-j.html>



Feature/Qualifier 対応一覧表: アルファベット順

Qualifier keys for source feature	
allele	altitude
anticodon	bio_material
artificial_location	cell_line
bound_moiety	cell_type
codon_start	chromosome
compare	clone
direction	clone_lib
EC_number	collected_by
estimated_length	collection_date
exception	country
experiment	cultivar
frequency	culture_collection
function	dev_stage
gap_type	ecotype
gene	environmental_sample
gene_synonym	focus
inference	germline
linkage_evidence	haplogroup
locus_tag	haplotype
mobile_element_type	host
mod_base	identified_by
ncRNA_class	isolate
note	isolation_source
number	lab_host
operator	lat_lon
PCR_conditions	macronucleus
product	map
pseudo	mating_type
pseudogene	mol_type
regulatory_class	note
replace	organelle
ribosomal_slippage	organism
rpt_family	PCR_primers
rpt_type	plasmid
rpt_unit_seq	proviral
satellite	rearranged
tag_peptide	segment
translation	serotype
transl_except	serovar
transl_table	sex
trans_splicing	specimen_voucher
	strain
	sub_clone
	sub_species
	sub_strain
	tissue_type
	transgenic
	variety

Feature keys	source	general
* assembly_gap	○	○
C region	○	○
* CDS	○	○
centromere	○	○
D-loop	○	○
D segment	○	○
* exon	○	○
gap	○	○
* intron	○	○
J segment	○	○
* LTR	○	○
* mat_peptide	○	○
misc_binding	○	○
misc_difference	○	○
* misc_feature	○	○
* misc_RNA	○	○
misc_structure	○	○
* mobile_element	○	○
modified_base	○	○
mRNA	○	○
* ncRNA	○	○
operon	○	○
oriT	○	○
precursor RNA	○	○
primer_bind	○	○
protein_bind	○	○
* regulatory	○	○
* repeat_region	○	○
rep_origin	○	○
* rRNA	○	○
* sig_peptide	○	○
* stem_loop	○	○
telomere	○	○
* tmRNA	○	○
transit_peptide	○	○
* tRNA	○	○
unsure	○	○
V region	○	○
V segment	○	○
* variation	○	○
* 3'UTR	○	○
* 5'UTR	○	○

必須  
使用可能  
使用不可  
脚注 1  
脚注 2  
脚注 3

- 注 1) gene\_synonym を記述する場合は、featureに gene または locus\_tag が必須です。  
注 2) exception, pseudo, pseudogene は同時に記述できません。exceptionを記述する場合はtranslationの記述が必須です。  
translationを記述する場合は、exceptionの記述が必須です。  
注 3) gap\_type が "within scaffold" または "repeat within scaffold" の場合、linkage\_evidence が必須です。

• DDBJ へ登録を推奨する feature と qualifier の組み合わせ

# 機能アノテーションの注釈: 例) CDS

Feature Key	CDS
Definition	coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes stop codon); feature includes amino acid conceptual translation.
Optional qualifiers	<code>/allele="text"</code> <code>/artificial_location="[artificial_location_value]"</code> <code>/citation=[number]</code> <code>/codon_start=&lt;1 or 2 or 3&gt;</code> <code>/db_xref="&lt;database&gt;:&lt;identifier&gt;"</code> <code>/EC_number="text"</code> <code>/exception="[exception_value]"</code> <code>/experiment="[CATEGORY:]text"</code> <code>/function="text"</code> <code>/gene="text"</code> <code>/gene_synonym="text"</code> <code>/inference="[CATEGORY:]TYPE[ (same species)][:EVIDENCE_BASIS]"</code> <code>/locus_tag="text" (single token)</code> <code>/map="text"</code> <code>/note="text"</code> <code>/number=unquoted text (single token)</code> <code>/old_locus_tag="text" (single token)</code> <code>/operon="text"</code> <code>/product="text"</code> <code>/protein_id="&lt;identifier&gt;"</code> <code>/pseudo</code> <code>/pseudogene="TYPE"</code> <code>/ribosomal_slippage</code> <code>/standard_name="text"</code> <code>/translation="text"</code> <code>/transl_except=(pos:&lt;base_range&gt;,aa:&lt;amino_acid&gt;)</code> <code>/transl_table=&lt;integer&gt;</code> <code>/trans_splicing</code>
Comment	<p>-----</p>

詳細は、The DDBJ/EMBL/GenBank Feature Table Definition  
[http://www.insdc.org/files/feature\\_table.html](http://www.insdc.org/files/feature_table.html)

# INSDC配列エントリ

LOCUS CP011474 2337451 bp DNA circular BCT 02-JUN-2015  
DEFINITION *Corynebacterium pseudotuberculosis* strain 12C, complete genome.  
ACCESSION CP011474  
VERSION CP011474.1 GI:827029750  
DBLINK BioProject: [PRJNA282769](#)  
BioSample: [SAMN03577816](#)  
KEYWORDS .  
SOURCE *Corynebacterium pseudotuberculosis*  
ORGANISM *Corynebacterium pseudotuberculosis*  
Bacteria; Actinobacteria; Corynebacteriales; Corynebacteriaceae;  
*Corynebacterium*.  
REFERENCE 1 (bases 1 to 2337451)  
AUTHORS Sousa,T.J., Mariano,D.C.B., Parise,D., Parise,M.T.D., Viana,M.V.C.,  
Guimaraes,L.C., Benevides,L.J., Rocha,F.S., Bagano,P., Almeida,S.  
and Azevedo,V.  
TITLE Direct Submission  
JOURNAL Submitted (15-MAY-2015) General Biology, Federal University of  
Minas Gerais, Av. Antonio Carlos 6627, Pampulha, Belo Horizonte,  
Minas Gerais 31270-901, Brazil  
COMMENT Source DNA and Bacteria are available from Laboratory of Cellular  
and Molecular Genetics, Federal University of Minas Gerais.

```
##Genome-Assembly-Data-START##  
Assembly Method :: Mira v. 3.9.18  
Reference-guided Assembly :: CP002924.1  
Genome Coverage :: 48x  
Sequencing Technology :: Ion Personal Genome Machine (PGMTM)  
System
```

```
##Genome-Assembly-Data-END##  
FEATURES Location/Qualifiers  
source 1..2337451  
/organism="Corynebacterium pseudotuberculosis"  
/mol_type="genomic DNA"  
/strain="12C"  
/isolation_source="abscess of sheep with CLA"  
/host="sheep"  
/db_xref="taxon:1719"  
/country="Brazil: Petrolina,PE"  
/lat_lon="9.39416 S 40.5096 W"  
/collection_date="2009"
```

```
gene 1..1812  
/gene="dnaA"  
/locus_tag="Cp12C_0001"  
CDS 1..1812  
/gene="dnaA"  
/locus_tag="Cp12C_0001"  
/codon_start=1  
/transl_table=11
```

## Structured Comment

# Structured comment

アノテーションファイル作成概説 | DDBJ

## Structured COMMENT (ST\_COMMENT)

- ST\_COMMENT は一定の記述ルールに従って、記載する COMMENT です。例えば、ゲノム配列（WGS も含む）における配列決定手法の概要（シーケンサ名、アセンブル方法、被覆度、他）、生物多様性関連のデータでサンプルの採取に関する概要などを記載することを想定しています。
- ST\_COMMENT は、外部のコンソーシアム、例えば、Barcode of Life Data Systems (BOLD)、Global Biodiversity Information Facility (GBIF)、Genomic Standards Consortium (GSC) など、の要請に基づいた記載内容、登録者が参加する研究コミュニティで規定された内容などを配列とともに登録し公開するために使用します。
- ST\_COMMENT では、COMMENT の内容と項目名、各項目の記載内容を規定します。
- Structured COMMENT の開始行では Qualifier に tagset\_id 、Value に COMMENT のタイトルを入力します。
- 規定に従って項目名を Qualifier として入力します。各項目に対応する具体的な内容を Value に入力します。
- ゲノムエントリ（WGS も含む）の登録では配列決定手法の実験的概要を示す "Genome-Assembly-Data" の記述を DDBJ は強く推奨しています。以下のように規定されています。
  - » tagset\_id は "Genome-Assembly-Data" です。
  - » Qualifier に入力する項目名は以下です。
    - Finishing Goal
    - Current Finishing Status
    - Assembly Method
    - Assembly Name
    - Genome Coverage
    - Sequencing Technology
  - "Assembly Method"、"Genome Coverage"、"Sequencing Technology" の 3 つは極力、入力して下さい。
  - 真核生物のゲノムの場合、"Assembly Name" を必ず入力してください。入力形式は "(organism の)属名+ 種名" もしくは "(organism の)一般名" + "数字" です（例：Btau\_4.0）。
  - "Finishing Goal" と "Current Finishing Status" では、入力可能な表記を制限しています。下記より選択して下さい。
    - » Standard Draft
    - » High-Quality Draft
    - » Improved High-Quality Draft
    - » Annotation-Directed Improvement
    - » Noncontiguous Finished
    - » Finished
  - 記載の可否や内容等については登録毎に個別に対応しますので、MSS の担当者にお問い合わせください。

# 【実習】Assembly Databaseを利用する

- <http://www.ncbi.nlm.nih.gov/assembly>
  - Browse by organismから”Cyanobacteria”のデータを検索する。興味の対象
  - Advancedから”Assembly level”=completeのデータを検索。配列Accessionなど様々な条件とその組み合わせ検索も可能です。”Show index list”を選択すると候補が表示されますので、色々試してみましょう。

# Assembly Database

<http://www.ncbi.nlm.nih.gov/assembly>

The screenshot shows the NCBI Assembly database homepage. At the top, there's a navigation bar with links for NCBI, Resources, How To, Home - Assembly - NCBI, fujisawa, My NCBI, and Sign Out. Below the navigation is a search bar with the word "Assembly" and a dropdown menu set to "Assembly". To the right of the search bar is a "Search" button. Underneath the search bar are links for "Advanced" and "Browse by organism". The main content area features a large image of a DNA sequence with red arrows pointing to specific regions. The word "Assembly" is prominently displayed in a large white font. Below it, the text reads "Genome assembly organization and additional information." On the left side, there's a sidebar titled "Using Assembly" with links to Assembly Help, Browse by Organism, NCBI Assembly Data Model, and Assembly Basics. In the center, there's a section titled "Submitting an Assembly" with links to Submission Information, Submission FAQ, AGP Specifications, and AGP Validation. On the right, there's a "Related Resources" section with links to Genome, Genome Reference Consortium, and Genome Remapping Service (Remap). At the bottom, there's a footer with links to various NCBI resources like PubMed, Bookshelf, and GenBank, as well as social media links for Facebook, Twitter, and YouTube. The footer also includes the National Library of Medicine logo and the USA.gov logo.

# Assembly Database

Organism browser – Assembly – NCBI

NCBI Resources How To

fujisawa My NCBI Sign Out

Assembly Assembly Search Advanced Browse by organism Help

Assembly information by organism eucaryotes (taxid:2759) Search by organism

Show only latest assemblies Show all assemblies

Items 1 - 25 of 2711 << First < Prev Page 1 of 109 Next > Last >>

Organism	Name	Submitter	Date	Genome representation	Assembly level	Version status	RefSeq category
Absidia idahoensis var. thermophila	<a href="#">Lramosa_hybrid_454_Illumina</a>	HKI JENA	08/05/2014	full	Scaffold	latest	representative genome
Acanthamoeba astronyxis	<a href="#">Acanthamoeba astronyxis</a>	UOL_CGR	01/07/2015	full	Contig	latest	representative genome
Acanthamoeba castellanii	<a href="#">Acas_2.0</a>	Baylor College of Medicine	03/22/2011	full	Scaffold	latest	na
Acanthamoeba castellanii	<a href="#">Acanthamoeba castellanii</a>	UOL_CGR	01/07/2015	full	Contig	latest	na
Acanthamoeba castellanii str. Neff	<a href="#">Acastellani.strNEFF v1</a>	JCVI	01/09/2013	full	Scaffold	latest	representative genome
Acanthamoeba culbertsoni	<a href="#">Acanthamoeba culbertsoni genome assembly</a>	UOL_CGR	01/06/2015	full	Contig	latest	representative genome
Acanthamoeba divionensis	<a href="#">Acanthamoeba divionensis</a>	UOL_CGR	01/07/2015	full	Contig	latest	representative genome
Acanthamoeba healyi	<a href="#">Acanthamoeba healyi</a>	UOL_CGR	01/07/2015	full	Contig	latest	representative genome
Acanthamoeba lenticulata	<a href="#">Acanthamoeba lenticulata</a>	UOL_CGR	01/07/2015	full	Contig	latest	representative genome
Acanthamoeba lugdunensis	<a href="#">Acanthamoeba lugdunensis</a>	UOL_CGR	01/07/2015	full	Contig	latest	representative genome
Acanthamoeba mauritaniensis	<a href="#">Acanthamoeba mauritaniensis</a>	UOL_CGR	01/07/2015	full	Contig	latest	representative genome
Acanthamoeba palestinensis	<a href="#">Acanthamoeba palestinensis</a>	UOL_CGR	01/07/2015	full	Contig	latest	representative genome
Acanthamoeba pearcei	<a href="#">Acanthamoeba pearcei</a>	UOL_CGR	01/07/2015	full	Contig	latest	representative genome

“cyanobacteria”を入力



# Assembly Database

Organism browser – Assembly – NCBI

NCBI Resources How To

fujisawa My NCBI Sign Out

Assembly Assembly Search Advanced Browse by organism Help

Assembly information by organism

blue-green algae (taxid:1117) Search by organism

Show only latest assemblies Show all assemblies

Organism	Name	Submitter	Date	Genome representation	Assembly level	Version status	RefSeq category
'Nostoc azollae' 0708	<a href="#">ASM19651v1</a>	US DOE Joint Genome Institute (JGI-PGF)	06/14/2010 full	Complete Genome	latest	representative genome	
[Oscillatoria] sp. PCC 6506	<a href="#">ASM18045v1</a>	ENSCP, Biochemistry, Paris, France	07/08/2010 full	Contig	latest	na	
[Scytonema hofmanni] UTEX 2349	<a href="#">ASM58268v1</a>	DOE Joint Genome Institute	02/27/2014 full	Scaffold	latest	representative genome	
Acaryochloris marina MBIC11017	<a href="#">ASM1810v1</a>	Washington University	10/16/2007 full	Complete Genome	latest	representative genome	
Acaryochloris sp. CCME 5410	<a href="#">ASM23877v2</a>	The Gordon and Betty Moore Foundation Marine Microbiology Initiative	06/03/2011 full	Contig	latest	na	
Anabaena cylindrica PCC 7122	<a href="#">ASM31769v1</a>	DOE Joint Genome Institute	12/07/2012 full	Complete Genome	latest	representative genome	
Anabaena sp. 90	<a href="#">ASM31270v1</a>	University of Helsinki	11/13/2012 full	Complete Genome	latest	na	
Anabaena sp. PCC 7108	<a href="#">ASM33213v1</a>	JGI	01/22/2013 full	Scaffold	latest	na	
Anabaena variabilis ATCC 29413	<a href="#">ASM20407v1</a>	DOE Joint Genome Institute	09/15/2005 full	Complete Genome	latest	representative genome	
Aphanizomenon flos-aquae 2012/KM1/D3	<a href="#">ASM78943v1</a>	Vilnius University	12/04/2014 full	Contig	latest	na	
Aphanizomenon flos-aquae NIES-81	<a href="#">Aflos454contigs199p</a>	Northern Illinois University	01/16/2014 full	Scaffold	latest	representative genome	
Aphanocapsa montana BDHKU210001	<a href="#">ASM81774v1</a>	Indian Institute Of Chemical biology	01/13/2015 full	Contig	latest	na	
Arthrospira maxima CS-328	<a href="#">ASM17355v1</a>	US DOE Joint Genome Institute (JGI-PGF)	10/15/2008 full	Contig	latest	na	

# Assembly Database

Advanced search – Assembly – NCBI

www.ncbi.nlm.nih.gov/assembly/advanced

NCBI Resources How To

fujisawa My NCBI Sign Out

Assembly Home Help

## Assembly Advanced Search Builder

Use the builder below to create your search

Edit Clear

### Builder

Assembly Level | Hide index list

chromosome (740)  
chromosome with gaps (567)  
complete genome (8935)  
contig (21713)  
gapless chromosome (7)  
scaffold (12629)

Previous 200  
Next 200

Refresh index

AND All Fields | Show index list

Search or Add to history

### History

Download history Clear history

Search	Add to builder	Query	Items found	Time
#22	Add	Search txid1148[Organism:exp] Sort by: SORTORDER	6	19:06:39

You are here: NCBI Write to the Help Desk

**GETTING STARTED**

- NCBI Education
- NCBI Help Manual
- NCBI Handbook
- Training & Tutorials

**RESOURCES**

- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps

**POPULAR**

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome

**FEATURED**

- Genetic Testing Registry
- PubMed Health
- GenBank
- Reference Sequences
- Gene Expression Omnibus
- Map Viewer
- Human Genome

**NCBI INFORMATION**

- About NCBI
- Research at NCBI
- NCBI News
- NCBI FTP Site
- NCBI on Facebook
- NCBI on Twitter
- NCBI on YouTube

# Assembly Database

Advanced search – Assembly – NCBI

www.ncbi.nlm.nih.gov/assembly/advanced

NCBI Resources How To

fujisawa My NCBI Sign Out

Assembly Home Help

## Assembly Advanced Search Builder

Use the builder below to create your search

Edit Clear

### Builder

Assembly Level

chromosome (740)  
chromosome with gaps (567)  
complete genome (8935)  
contig (21713)  
gapless chromosome (7)  
scaffold (12629)

Previous 200  
Next 200

Refresh index

AND All Fields

Show index list

Search or Add to history

### History

Download history Clear history

Search	Add to builder	Query	Items found	Time
#22	Add	Search txid1148[Organism:exp] Sort by: SORTORDER	6	19:06:39

You are here: NCBI Write to the Help Desk

**GETTING STARTED**

- NCBI Education
- NCBI Help Manual
- NCBI Handbook
- Training & Tutorials

**RESOURCES**

- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps

**POPULAR**

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome

**FEATURED**

- Genetic Testing Registry
- PubMed Health
- GenBank
- Reference Sequences
- Gene Expression Omnibus
- Map Viewer
- Human Genome

**NCBI INFORMATION**

- About NCBI
- Research at NCBI
- NCBI News
- NCBI FTP Site
- NCBI on Facebook
- NCBI on Twitter
- NCBI on YouTube

# Assembly Database

Advanced search – Assembly – NCBI

www.ncbi.nlm.nih.gov/assembly/advanced

NCBI Resources How To

fujisawa My NCBI Sign Out

Assembly Home Help

### Assembly Advanced Search Builder

"nc 000962 2"[Reference Guided Assembly]

Edit Clear

#### Builder

Reference Guided As ▾ "nc 000962 2"[Reference Guided Assembly] Hide index list

lactobacillus plantarum wcfs1 (1)  
lysinibacillus sphaericus c3 41 (3)  
mshr1153 (1)  
mycobacterium bovis af2122 97 (12)  
mycobacterium tuberculosis h37rv (2)  
**nc 000962 2 (10)**  
nc 000962 3 (4)  
nc 000964 3 (4)  
nc 002488 3 (1)  
nc 003143 1 (8)

Previous 200  
Next 200

Refresh index Show index list

AND All Fields

Search or Add to history

---

#### History

Download history Clear history

Search	Add to builder	Query	Items found	Time
#22	Add	Search txid1148[Organism:exp] Sort by: SORTORDER	6	19:06:39

You are here: NCBI Write to the Help Desk

**GETTING STARTED**

NCBI Education  
NCBI Help Manual  
NCBI Handbook  
Training & Tutorials

**RESOURCES**

Chemicals & Bioassays  
Data & Software  
DNA & RNA  
Domains & Structures  
Genes & Expression  
Genetics & Medicine  
Genomes & Maps

**POPULAR**

PubMed  
Bookshelf  
PubMed Central  
PubMed Health  
BLAST  
Nucleotide  
Genome

**FEATURED**

Genetic Testing Registry  
PubMed Health  
GenBank  
Reference Sequences  
Gene Expression Omnibus  
Map Viewer  
Human Genome

**NCBI INFORMATION**

About NCBI  
Research at NCBI  
NCBI News  
NCBI FTP Site  
NCBI on Facebook  
NCBI on Twitter  
NCBI on YouTube

# Minimum standards for annotating complete genomes

---

[http://www.ncbi.nlm.nih.gov/genome/annotation\\_prok/standards](http://www.ncbi.nlm.nih.gov/genome/annotation_prok/standards)

## Minimum standards for annotating complete genomes

### 1. ANNOTATION SHOULD FOLLOW INSDC SUBMISSION GUIDELINES (GenBank/ENA/DDBJ)

- a. Prior to genome submission a submitted Bioproject record with a registered locus\_tag prefix is required according to accepted guidelines

<http://www.ncbi.nlm.nih.gov/genomes/locustag/Proposal.pdf>

- b. The genome submission should be valid according to feature table documentation

[http://insdc.org/documents/feature\\_table.html](http://insdc.org/documents/feature_table.html)

### 2. MINIMAL GENOME ANNOTATION SHOULD HAVE

- a. At least one copy of rRNAs (5S, 16S, 23S) of appropriate length and corresponding genes with locus\_tags

- b. At least one copy of tRNAs for each amino acid and corresponding genes with locus\_tags

- c. Protein-coding genes with locus\_tags (see below) and corresponding CDS

### 3. VALIDATION CHECKS AND ANNOTATION MEASURES

Validation checks should be done prior to the submission. NCBI has already provided numerous tools to validate and ensure correctness of annotation. Additional checks will be put in place to ensure the minimal standards are met.

Statistical measures that are used for annotation quality assessment include:

- a. Feature counts by feature type

- b. Protein coding gene count vs genome size ratio

- c. Percent of short (<30 aa) proteins

- d. Percent of coding regions with a standard start codon

- e. Count of protein coding regions with “hypothetical protein” product

### 4. EXCEPTIONS

Exceptions (unusual annotations, annotations not within expected ranges) should be documented and strong supporting (experimental) evidence should be provided.

# 機能アノテーションガイドライン

---

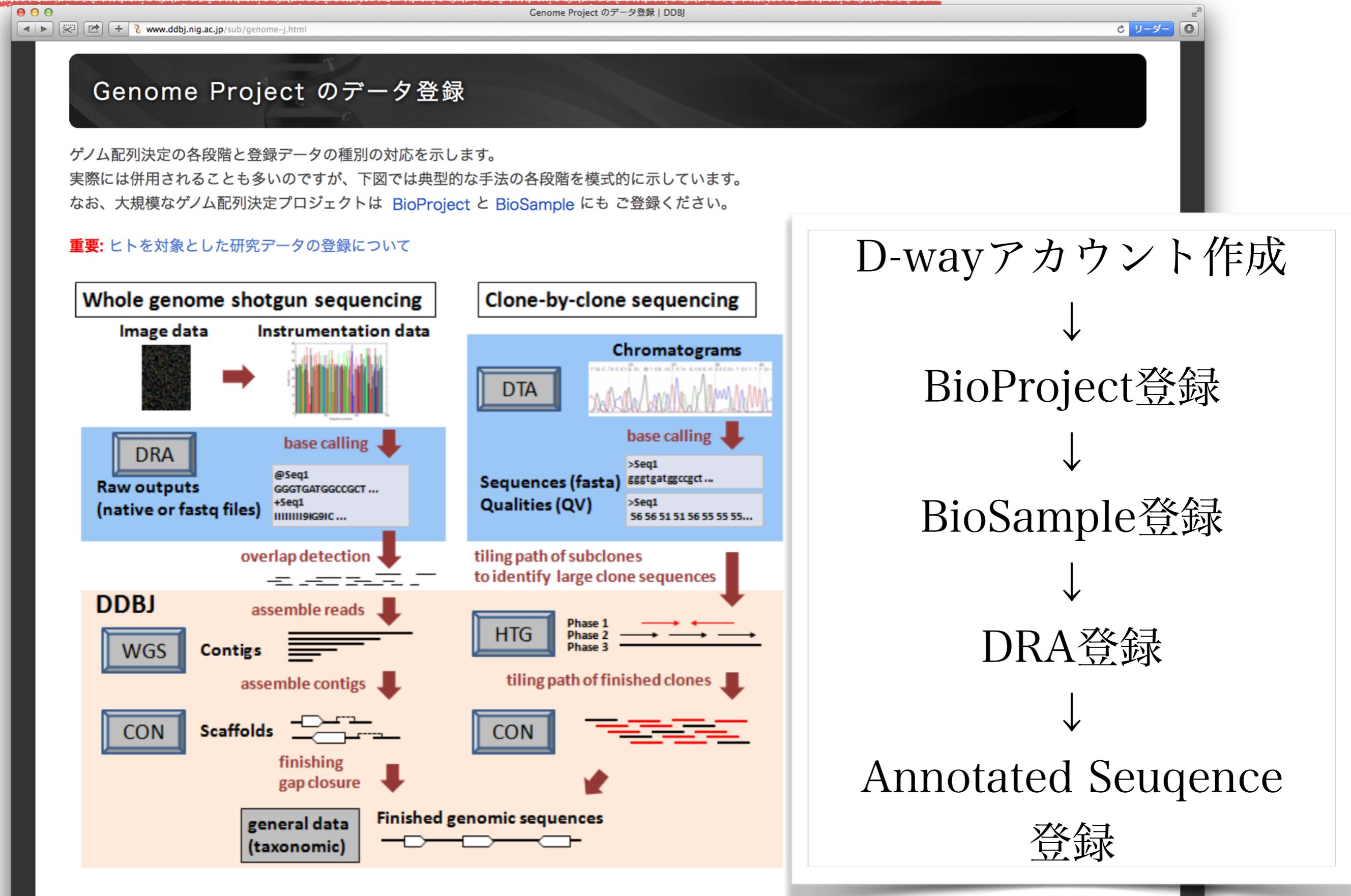
- UniProtKB/Swiss-Prot
  - Protein naming guidelines
    - <http://www.uniprot.org/docs/nameprot>
  - Protein nomenclature publication list
    - <http://www.uniprot.org/docs/nomlist>
  - Prokaryotic protein naming guidelines
    - <http://www.uniprot.org/docs/proknameprot>

## 参考) ゲノムアノテーションにおける原核生物と真核生物の違い

---

- 原核生物
  - 遺伝子領域の予測精度が高い
- 真核生物
  - exon-intron構造を持ち、予測精度が高くない（60-70%）ため、ESTやfull-length cDNAなどのデータを用いたマニュアルキュレーションが必要
- 参考) ゲノムアノテーション～真核生物を中心に～
  - <https://www.youtube.com/watch?v=1Y0OdHEBYK4>

# DDBJへのデータ登録



ここまでで前提知識です。

!?

# 微生物ゲノムアノテーションパイプラインの紹介

---

- MiGAP
  - <http://www.migap.org>
- RAST
  - <http://rast.nmpdr.org>
- NCBI PGAP
  - [http://www.ncbi.nlm.nih.gov/genome/annotation\\_prok/](http://www.ncbi.nlm.nih.gov/genome/annotation_prok/)
  - prokka (環境構築&コマンドライン)
    - <http://www.vicbioinformatics.com/software.prokka.shtml>

# MiGAP: 入力

MiGap

migap.ddbj.nig.ac.jp/mgap/jsp/index.jsp

**MiGAP Microbial Genome Annotation Pipeline ver2.19**

Logout Help Contact Us

LDAP\_tafujisa (b-MiGAP) 2015/06/17 08:42:39 [View Menu] [Hide Menu]

Pipe Line [Running:8 Waiting:2]

Pipe Line Name:

Upload Filename:  ファイル未選択  
or paste data in box below. ([Sample data](#))

or paste data in box below. ([Sample data](#))

Linear  Circular  Bacteria  Archaea  Eukarya \*FUNGI\* transl\_table: 11

**Run** **Clear**

**Change User Level**

b-MiGAP  
 s-MiGAP  
 g-MiGAP

**Set**

# MiGAP: 出力

MiGap

migap.ddbj.nig.ac.jp/mgap/jsp/index.jsp

C リーダー

**MiGAP Microbial Genome Annotation Pipeline ver2.19**

Logout Help Contact Us LDAP\_tafujisa (b-MiGAP) 2015/06/17 08:42:39 [View Menu] [Hide Menu]

Pipe Line History

Pipe Line History Change User Level Current Process

LDAP\_tafujisa | Hidden Data List Contig

<< < [1/1] > >> << < [1/1] > >> contig

ajacs54\_demo(whole)  
Leptolyngbya\_boryana\_dg5(whole)  
Leptolyngbya\_boryana\_wt(whole)  
Synechocystis No. 3(whole)  
Bacillus subtilis 168  
MG1655

**Basic Information**

Filename: nc\_000913.3.fasta  
Contig: 1  
Total Length: 4641652  
CDS: 4319  
RBS: 4241  
rRNA: 22  
tRNA: 88  
Run: 2014/02/20 14:57:46

**ORF & RNA Extract**

Start: 2014/02/20 14:57:46  
End: 2014/02/20 15:04:31  
Software: MetaGeneAnnotator 1.0  
tRNAscan-SE 1.23  
NCBI BLAST 2.2.18  
RNAmmer 1.2

**Annotation**

Start: 2014/02/20 15:04:31  
End: 2014/02/20 16:29:43  
Software: NCBI BLAST 2.2.18  
DB: COG[20030417]:4292  
RefSeq release56[20130314]:4290  
TrEMBL release2013\_03[20130306]:28

**Download**

Log File [pipeline.log](#)  
N.A.: [result-na.fasta.tar.gz](#)  
A.A.: [result-aa.fasta.tar.gz](#)

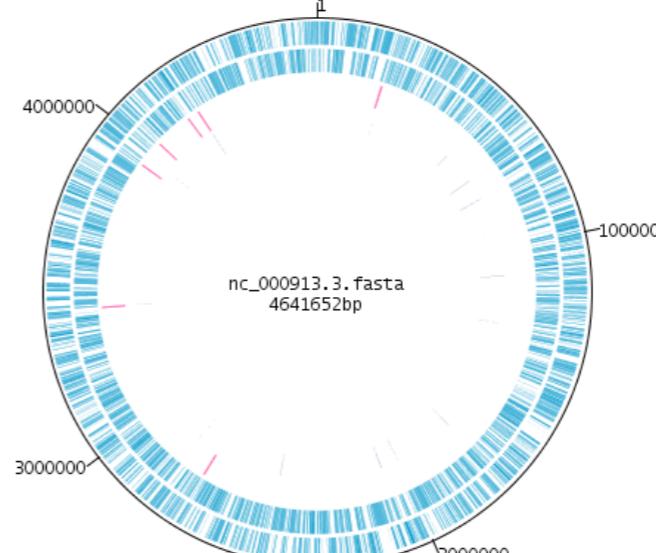
**Before Annotation**

CSV: [result.csv.tar.gz](#)  
Genbank: [result.gbk.tar.gz](#)  
EMBL: [result.embl.tar.gz](#)  
DDBJ: [result.fasta.tar.gz](#)  
[result.annt.tar.gz](#)  
[result.ddbj.tar.gz](#)  
[result.ddbj.tar.gz\(multi\)](#)

**After Annotation**

CSV: [result-a.csv.tar.gz](#)  
Genbank: [result-a.gbk.tar.gz](#)  
EMBL: [result-a.embl.tar.gz](#)  
DDBJ: [result-a.fasta.tar.gz](#)  
[result-a.annt.tar.gz](#)  
[result-a.ddbj.tar.gz](#)  
[result-a.ddbj.tar.gz\(multi\)](#)

GFF: [result-off.tar.gz](#)



# ゲノムアノテーションの精度向上

- GenomeMatcher (MacOSX上のみ)
  - <http://www.ige.tohoku.ac.jp/joho/gmProject/gmhomeJP.html>
  - Comparative Sequences/アノテーション支援
    - <http://www.ige.tohoku.ac.jp/joho/gmProject/manualJP/AnnotationSupport.html>
    - ORFのblastp解析によるStart positionの修正
    - ORFの抽出と消去

# ゲノムアノテーションの精度向上

<http://genome.annotation.jp/genomerefine/>

- GenomeRefineの紹介

The screenshot shows the 'Genome Refine' web application. The title bar says 'Genome Refine'. The main content area has a grey header 'Genome Refine'. Below it, there are several sections with headings: 'About', 'Menu', 'Functions and Features', 'Sementic WEB technology', 'presentations and tutorials', and 'Collaboration Services and Tools'. Each section contains descriptive text and links.

**About**  
Genome Refine は微生物ゲノムアノテーションの高品質化および解析支援のためのウェブサービスです。  
現在、限定α版として公開しています。

**Menu**  
• アカウント作成

**Functions and Features**

- ユーザ・グループ認証機能**  
OpenIDユーザ認証拡張によりグループ情報をもつ認証機能。参加グループのみでゲノムデータ共有され、公開・非公開が選択可能です。TogoAnnotationなどの連携サービスも同じアカウントで利用出来ます。
- ゲノムデータ入力・編集機能**  
微生物ゲノムアノテーション実行パイプラインMiGAPの結果(fasta形式およびcsv形式)を入力します。
- メタデータ入力・編集機能**  
GenBank形式ファイルおよび比較解析に必要な情報をユーザが入力します。Organismは入力必須項目です。
- データ形式変換機能**  
Resource Description Framework (RDF)形式および標準的なGenBank, GFF3, FASTAで出力します。編集されたメタデータ、アノテーションが反映されます。
- 自動ゲノムアノテーション編集機能 (\*)**  
NITE標準アノテーション辞書とテキストマイニング技術を用いた遺伝子産物名称のアノテーション高精度化ツールTogoAnnotatorを利用。自動によるアノテーションを高度化を実施します。
- ゲノムデータベース機能 (\*)**  
CyanoBaseと同じデータベースシステムによるゲノムデータベースを提供。キーワード検索およびBlast検索が実行可能です。
- 手動ゲノムアノテーション編集機能 (\*)**  
CyanoBaseと同じユーザインターフェースをもつデータベース上で手動アノテーション編集機能を提供。遺伝子ページから編集が可能で、データはTogoAnnotationに保存されます。

\* 開発中のため機能を制限しております。本機能を利用してみたい方は個別にご相談下さい。

**Sementic WEB technology**

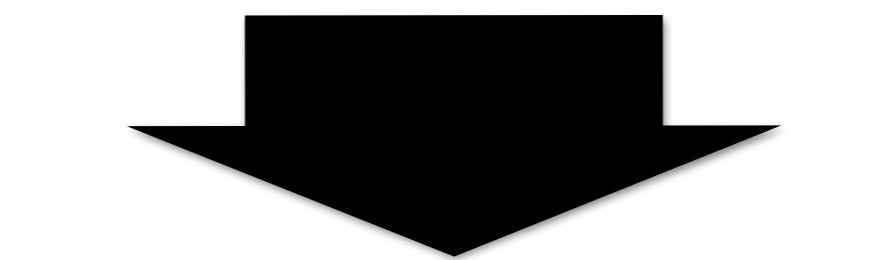
- ゲノム情報RDF**  
DBCLS,DDBJで連携して開発しているゲノム情報RDFを生成
- メタデータOWL**  
入力メタデータ項目についてはOWLで定義

**presentations and tutorials**

- トーゴーの日シンポジウム2013

**Collaboration Services and Tools**

# 増大するNGSデータにおける課題



NGSデータ

ゲノム情報

# 微生物ゲノムアノテーションデータフローとDDBJ関連サービス

NGS sequence read



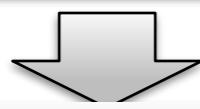
Genome sequence



Genome structural  
annotation



Genome functional  
annotation



Genome annotation  
refinement



Data/Database  
publishing

## DDBJ Read Annotation Pipeline

NGSデータを入力として配列アセンブルを行なう自動パイプライン、Galaxyによる解析ワークフローも提供



微生物ゲノム配列に対して遺伝子予測と簡易機能アノテーションを行う自動パイプライン



- Data curation
- Database construction
- INSDC/DDBJ submission

データフローのボトルネック

ゲノムアノテーション高品質化プロセスのスループット向上の必要性

Twitter / synobu: 要望 : MiGAPから直接DDBJに配列登録できるようにしてく ...

Twitter, Inc. [twitter.com/synobu/status/224077599700299780](https://twitter.com/synobu/status/224077599700299780)

リーダー

Trac1w TracMy H25TGAQ1 insdc-group:...tation\_rule AnnotateIt (99+) annot...ogle Groups Evernote A3-2:SCO0190

Twitter / synobu: 要望 : MiGAPから直接DDBJに配列登録できるようにしてく ...

ホーム つながり 見つける アカウント 検索

 so  
@synobu

 フォロー中

要望 : MiGAPから直接DDBJに配列登録できるようにしてください！MiGAPの連続投げの時間制限を解除してください！ #cyanoinfo

返信 リツイート ★ お気に入りに登録 その他

2012年7月14日 - 18:47

@synobuさんへ返信する

 Osamu Ogasawara @oogasawa 7月14日  
@synobu 技術的なことは大山さんに相談されてはいかがでし。

# 微生物ゲノムアノテーションデータフローとDDBJ関連サービス

NGS sequence read



Genome sequence



Genome structural  
annotation



Genome functional  
annotation



Genome annotation  
refinement



Data/Database  
publishing

## DDBJ Read Annotation Pipeline

NGSデータを入力として配列アセンブルを行なう自動パイプライン、Galaxyによる解析ワークフローも提供



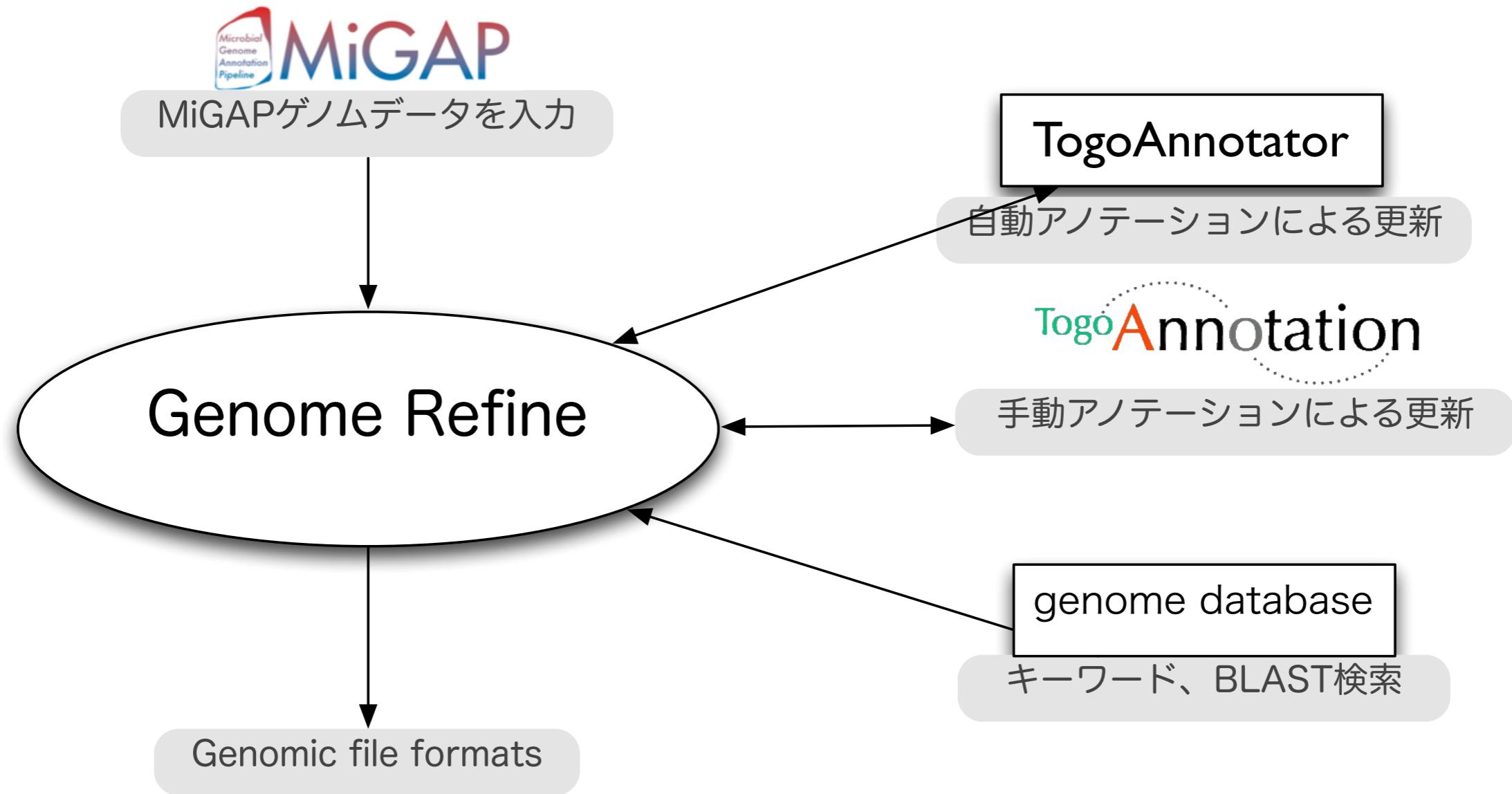
微生物ゲノム配列に対して遺伝子予測と簡易機能アノテーションを行う自動パイプライン

## Genome Refine

ゲノムアノテーション精度向上のための  
ウェブサービス開発

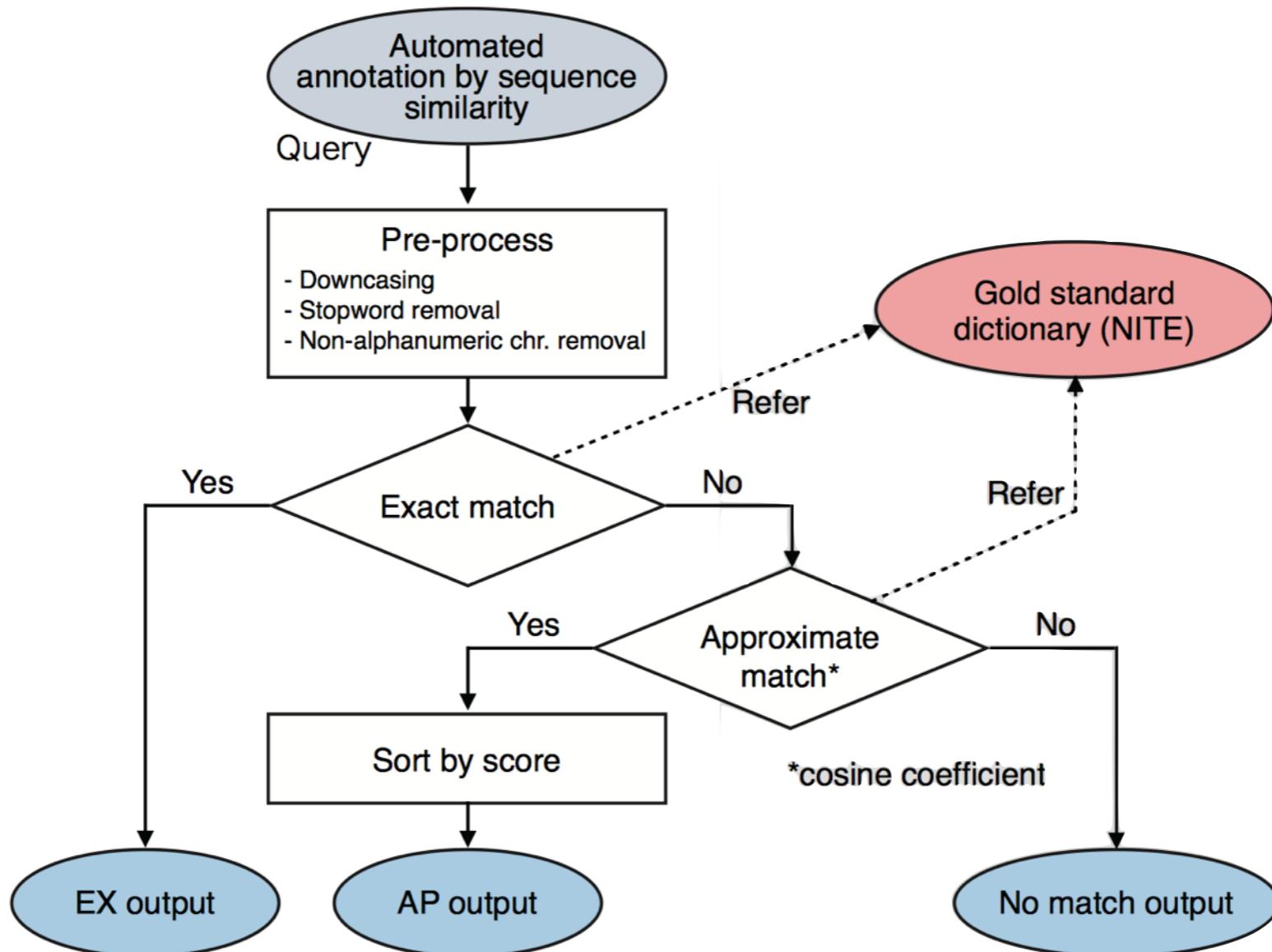
# Genome Refineとは？

未公開微生物ゲノムデータを低成本にアノテーションしてアップロードするためのウェブサービス



# TogoAnnotatorとは？

テキスト処理技術を用いた遺伝子機能アノテーションの自動リファインメントツール



Schematic diagram showing the process of product name assignment via TogoAnnotator

# 例) TogoAnnotatorによる自動リファインメント

テキスト処理技術を用いた遺伝子機能アノテーションの自動リファインメントツール

## Examples of TogoAnnotator outputs

Gene	Automated annotation pipeline	TogoAnnotator	
GOHSU_40_00070	RNA polymerase ECF <b>subfamily</b> sigma factor	RNA polymerase ECF-type sigma factor	EX
MA3A0122S_4041	RNA polymerase ECF- <b>subfamily</b> sigma factor	RNA polymerase ECF-type sigma factor	EX
GOHSU_24_00170	DNA gyrase <b>subunit A</b>	DNA gyrase subunit A	EX
MA3A0122S_5168	DNA gyrase, <b>A subunit</b>	DNA gyrase subunit A	AP
MA3A0122S_4747	<b>putative</b> GntR transcriptional <b>regulatory protein</b>	GntR family transcriptional regulator	AP
MA3A0122S_1388	transcriptional regulator, <b>GntR family</b>	GntR family transcriptional regulator	AP
MA3A0122S_0916	transcriptional regulator, <b>GntR family domain protein</b>	GntR family transcriptional regulator	AP
GOHSU_04_00020	<b>putative</b> GntR family transcriptional regulator	GntR family transcriptional regulator	EX
MA3A0122S_1880	<b>gntR-family</b> transcriptional regulator	GntR family transcriptional regulator	EX
MA3A0122S_2499	<b>ABC D-methionine transporter, permease component</b>	methionine ABC transporter permease protein	AP
MA3A0122S_0906	<b>putative DTDP-GLUCOSE 4,6-DEHYDRATASE RMLB2</b>	dTDP-glucose 4,6-dehydratase	AP
MA3A0122S_1010	ATP synthase <b>F0, A subunit</b>	ATP synthase subunit a	AP



Licensed under a Creative Commons 表示 2.1 日本 license (c)2013

山本泰智 (DBCLS), 岡本忍 (DBCLS), 宮沢せいは (NITE), 市川夏子 (NITE), 藤田信之 (NITE)

# TogoAnnotationとは？

ソーシャルブックマークの仕組みを用いて生物エンティティに特化したアノテーションを行なうウェブサービス。URLに対してタグとして遺伝子名称をアノテーションを実施

The screenshot shows the TogoAnnotation website interface. On the left, there's a sidebar with links to 'About TogoAnnotation', 'News' (including a tweet from 'togo\_annotation' about a genome annotation of Bradyrhizobium sp. S23321), and 'Recent Annotations' (listing entries like Ann:TG:4742 and Ann:TG:2095). The main content area has a header 'Social Genome Annotation - TogoAnnotation'. It displays a search bar, a navigation menu with 'TogoAnnotation', '統合アノテーション > Ann:TG:4742', and user links 'fujisawa | アカウント設定 | インポート/エクスポート | サインアウト'. Below the header, there's a large green bar with the text 'Ann:TG:4742 編集'. The main panel contains several tables and lists related to the gene AdpA, including:

- Gene names (5):** AdpA (708 annotations)
- PubMed IDs (55):** A grid of 55 PMID numbers ranging from 22449632 to 17635544.
- PPI (0) :** -
- Experiments (0) :** -
- ユーザ (5) :** hinamana, keikeikeiko, kkato, syamada, eishii
- アノテーション:** 2600 エントリ [SearchView | TableView | GeneIndexingView | IndexedRefView | ReferenceView]
- タグ (127) :** GA, pname:AdpA, results, discussion, introduction, body, materials and methods, abstract, experimental procedures, fig3, fig1, fig4, fig5, fig2, fig6, pname:transcriptional activator, pname:A-factor-dependent protein, pname:A-factor-responsive protein, pname:AdpAg, supplemental materials and methods, pname:transcriptional factor, pname:A-factor-dependent transcriptional activator, pname:A-factor-responsive binding protein, figS5, pname:transcriptional regulator, pname:global transcriptional regulator for the onset of morphological and physiological development, pname:A-factor-dependent regulator, pname:AraC/XylS family transcriptional regulator, pname:curated, pname:The AraC/XylS family transcriptional regulator, pname:A-factor-responsive transcriptional activator, pname:A-factor-dependent transcriptional activator, pname:transcriptional activator protein, tableS3, pname:transcriptional regulator, acknowledgement, Concluding remarks, tableS1, pname:the global transcriptional regulator for the onset of morphological and physiological development, pname:AraC/XylS transcriptional regulator, pname:AraC-like protein, tableS1, tableS4, pname:A-factor-dependent transcriptional regulator, GG, pname:A-factor-dependent activator protein, gene:SGR4742, pname:hypothetical protein, pname:A transcriptional activator, table3, pname:ArpA.
- 編集:** アノテーションを編集

At the bottom, there are two examples of annotations for Streptomyces griseus NBRC 13350 SGR\_4742:

- keikeikeiko GA introduction pmid:22133433 pname:AdpA**: AdpA orthologs acting as positive regulators for antibiotic biosynthesis have been found in several Streptomyces. (2013-02-12)
- keikeikeiko GA introduction pmid:22133433 pname:AdpA**: AdpA is active in the A-factor regulatory cascade, and switches on a number of genes required for both morphological development and secondary-metabolite formation [6,7]. (2013-02-12)

<http://togo.annotation.jp>

例) Streptomyces griseus NBRC13350 SGR\_4742 遺伝子

# TogoAnnotationを用いた手動アノテーション

The screenshot shows the TogoAnnotation GeneView interface for gene GR3\_10060. The top navigation bar includes a search bar with 'search in Gene symbols, Gene products' and a 'save' button. Two red arrows point to the 'cancel' button in the gene symbol section and the 'edit' button in the gene product section.

**GeneView: GR3\_10060**

**GFF3**

Summary | Genomic context | Transcript | Peptide | Protein domains | Mutants | Annotations | References | Homology | Ortholog table | External links | API

Summary | ? | +/- | Top

ID	GR3_10060
Type	protein_coding_gene
Gene symbol	proS, proS => proS <input type="text" value="proS"/> <span>save</span> <span>cancel</span>
	proS, proS
Gene symbol Extracted from literature	
Gene product	proline-tRNA ligase <span>edit</span>
Gene product Extracted from literature	
Functional category	

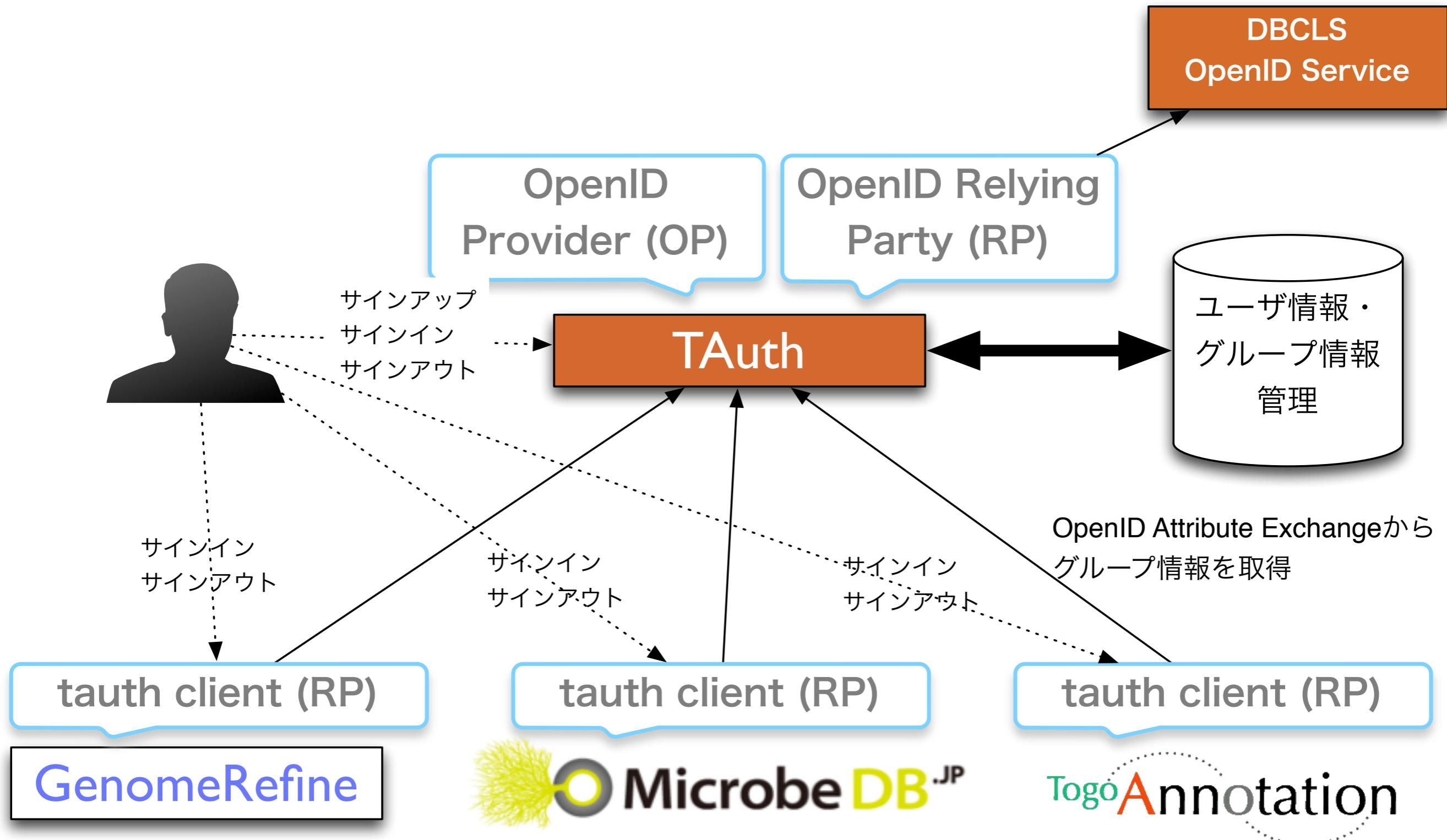
**Genomic context** | ? | +/- | Top

Location	sequence1:5840..7642
Strand	direct

Error: undefined; undefined

編集されたアノテーションデータはTogoAnnotationにブックマークデータとして保存され、ゲノムデータに上書きされる

# TAuth: ユーザ・グループ情報管理システム



本認証システムの拡充によって、アクセスレベルの制限システムの構築

# Genome Refineの使い方

## ゲノム情報を入力・更新する

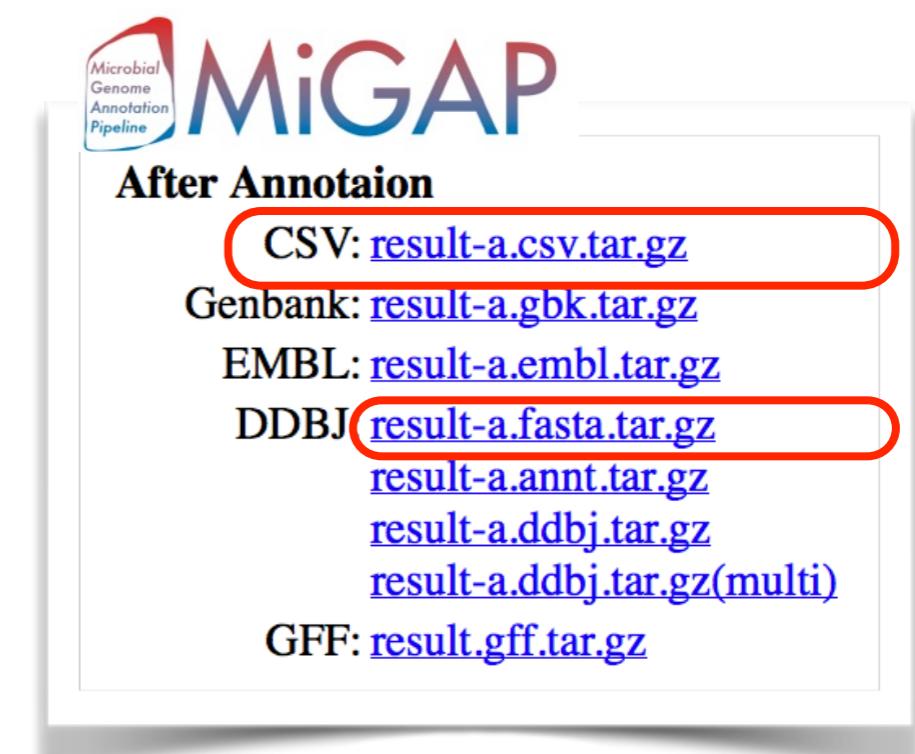
The screenshot shows the 'Project' section of the Genome Refine interface. It includes fields for 'Create new group or select your group\*' (set to 'sample'), 'Assign access control levels of database and genomic data files' (set to 'Private'), and 'Genome Analysis' options like 'Create genome database' and 'Improvement of gene annotation using TogoAnnotator'. The 'Genome Sequence' section shows 'Sequence files\*' and 'Annotation files\*' both set to 'result-a.fasta.tar.gz'. The 'Project Metadata' section contains various input fields for assembly method, name, coverage, isolation source, locus tag prefix, project description, sequence technology, strain, organism (set to 'Escherichia coli MG1655'), and sequences. An 'Execute' button is at the bottom.

メタデータを入力  
(Organism名は必須)

共有グループを作成or選択

公開 or 非公開を選択

MiGAP解析結果の圧縮ファイル  
形式をそのまま入力



約10分程度でデータベースが作成されます。

# 任意のグループに対して非公開ゲノムデータベースを提供

## グループ管理ページの一例

The screenshot displays three windows illustrating a group management system for a genome database.

**Top Window:** A browser window titled "Tconf" showing a "sample Project List". The table lists two projects: GR23 and GR24. Both entries show "Bacillus subtilis subsp. subtilis 168" as the organism, version 3, created and updated on 2014-03-06, and status "processed". The "Genome Database" column contains a checkmark, and the "Metagenome Analysis" column is empty. The "Project" column shows "edit delete download" links. A "Sign out" button is at the top right.

**Middle Window:** A browser window titled "241/GR24 – Top" showing the "Admin" section. It includes a "UserAccount" sidebar and a main area with tabs for "Action", "Project List", and "Table of Contents". The "Action" tab is active, showing a "GenomeDatabase" section with a link to "http://genome.microbedb.jp/241". A large arrow points from the "Project List" link in the sidebar to the "Project List" tab in the main area.

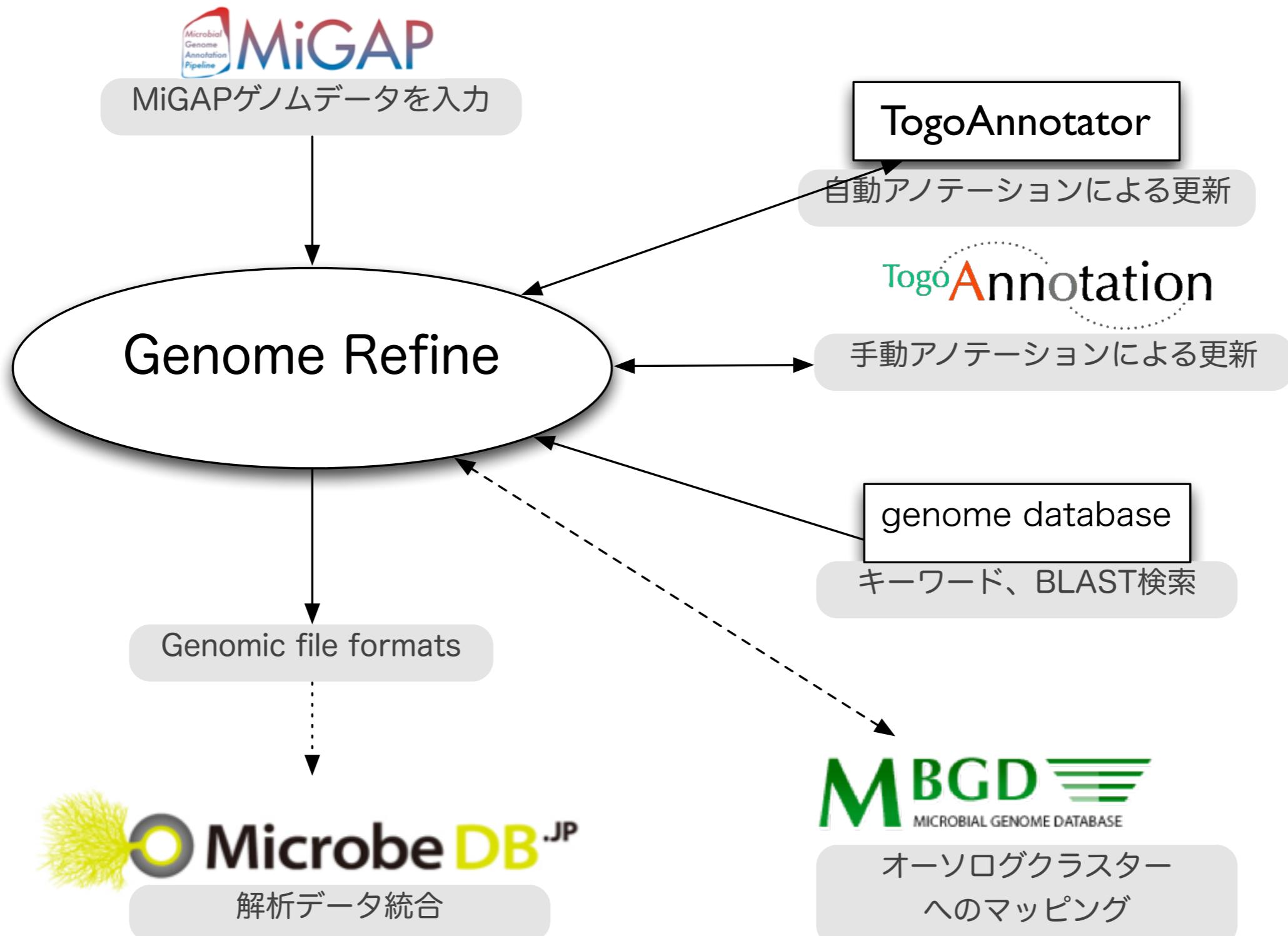
**Bottom Window:** A browser window titled "auth.annotation.jp" showing a "sample Users" page. It lists a user "tf@nig.ac.jp" and has a "Update Group" button at the bottom. The URL is "https://auth.annotation.jp/group\_admin/groups/241/edit".

# 課題と現在の取り組み

---

- 自動アノテーション精度向上
  - INSDC登録用TogoAnnotator辞書開発
  - DDBJ登録ガイドラインの策定
- DDBJ登録形式ファイルの生成
  - Submitter, BioProject, BioSample情報の入力項目拡張
  - ドラフトゲノムCON/WGS登録AGP形式ファイル生成
  - 真核微生物データの対応
- 解析パイプライン連携強化
  - ユーザ利便性向上を目的としたMiGAP連携
  - メタゲノム解析パイプラインMeGAP連携

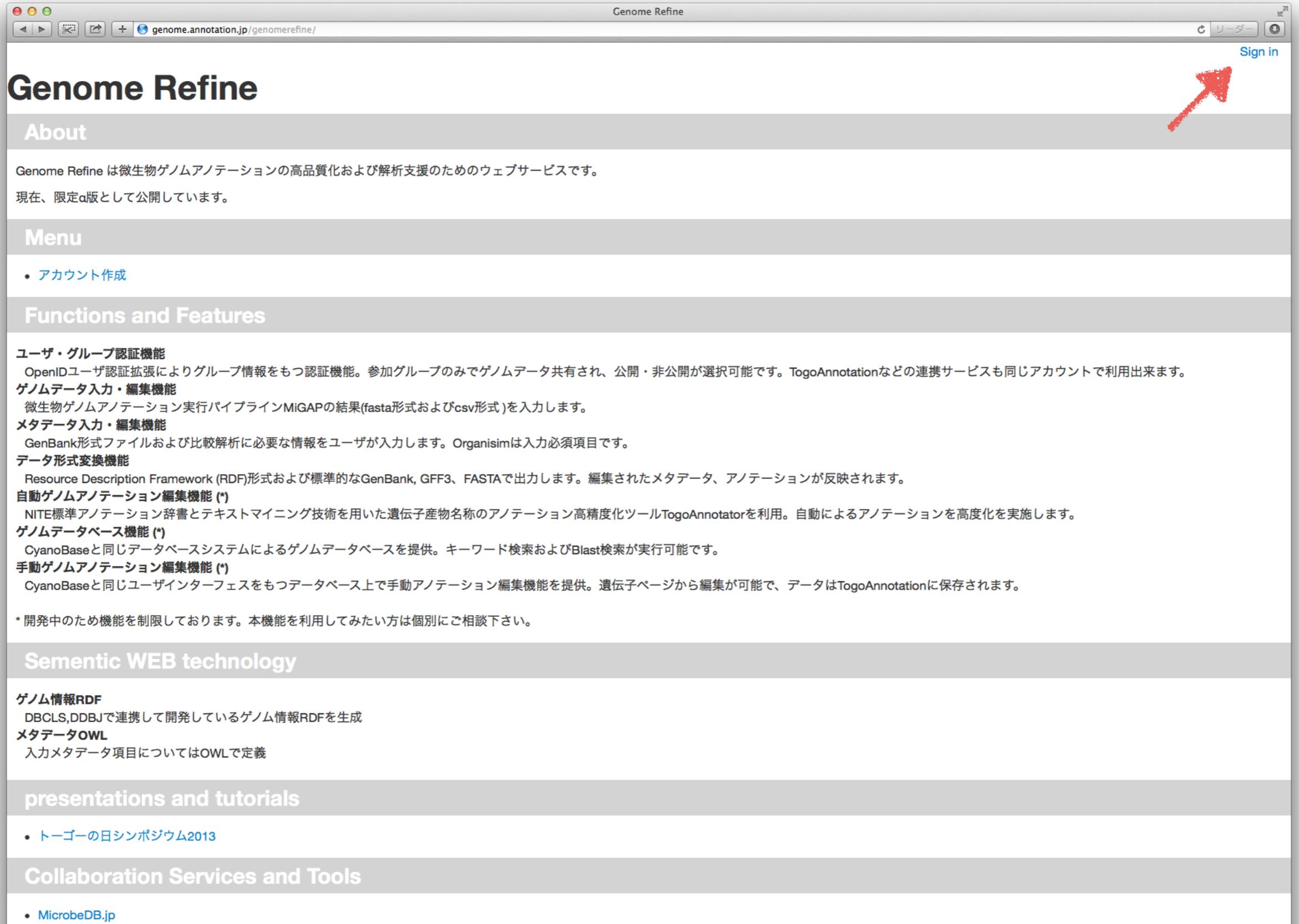
# MicrobeDB.jpプロジェクト



微生物ゲノムアノテーションの高品質化および解析支援のためのウェブサービス

# 【発展】 GenomeRefineを利用して見る

<http://genome.annotation.jp/genomerefine/>



The screenshot shows the 'Genome Refine' web application. At the top, there's a navigation bar with icons for back, forward, and search, followed by the URL 'genome.annotation.jp/genomerefine/'. On the far right of the bar is a 'Sign in' link. Below the bar, the main title 'Genome Refine' is displayed. A red arrow points from the bottom right towards the 'Sign in' link. The page content includes sections for 'About', 'Menu' (with a single item 'アカウント作成'), 'Functions and Features' (listing various features like User-Group Authentication, Genome Data Input/Editing, etc.), 'Sementic WEB technology' (mentioning RDF and OWL), 'presentations and tutorials' (listing 'トーゴーの日シンポジウム2013'), and 'Collaboration Services and Tools' (listing 'MicrobeDB.jp').

Genome Refine は微生物ゲノムアノテーションの高品質化および解析支援のためのウェブサービスです。

現在、限定α版として公開しています。

**Menu**

- アカウント作成

**Functions and Features**

ユーザ・グループ認証機能  
OpenIDユーザ認証拡張によりグループ情報をもつ認証機能。参加グループのみでゲノムデータ共有され、公開・非公開が選択可能です。TogoAnnotationなどの連携サービスも同じアカウントで利用出来ます。

ゲノムデータ入力・編集機能  
微生物ゲノムアノテーション実行パイプラインMiGAPの結果(fasta形式およびcsv形式)を入力します。

メタデータ入力・編集機能  
GenBank形式ファイルおよび比較解析に必要な情報をユーザが入力します。Organismは入力必須項目です。

データ形式変換機能  
Resource Description Framework (RDF)形式および標準的なGenBank, GFF3, FASTAで出力します。編集されたメタデータ、アノテーションが反映されます。

自動ゲノムアノテーション編集機能 (\*)  
NITE標準アノテーション辞書とテキストマイニング技術を用いた遺伝子産物名称のアノテーション高精度化ツールTogoAnnotatorを利用。自動によるアノテーションを高度化を実施します。

ゲノムデータベース機能 (\*)  
CyanoBaseと同じデータベースシステムによるゲノムデータベースを提供。キーワード検索およびBlast検索が実行可能です。

手動ゲノムアノテーション編集機能 (\*)  
CyanoBaseと同じユーザインターフェスをもつデータベース上で手動アノテーション編集機能を提供。遺伝子ページから編集が可能で、データはTogoAnnotationに保存されます。

\* 開発中のため機能を制限しております。本機能を利用してみたい方は個別にご相談下さい。

**Sementic WEB technology**

ゲノム情報RDF  
DBCLS,DDBJで連携して開発しているゲノム情報RDFを生成

メタデータOWL  
入力メタデータ項目についてはOWLで定義

**presentations and tutorials**

- トーゴーの日シンポジウム2013

**Collaboration Services and Tools**

- MicrobeDB.jp

# Genome Refineにアクセスする

---

- demo@annotation.jp
- パスワード: ajacs54

# GenomeRefineを利用して見る

auth.annotation.jp

auth.annotation.jp

**Sign in**

**Sign in with Email and Password**

Email  
demo@annotation.jp

Password  
.....

Remember me

Sign in

**Sign in with OpenID**

Openid url  


Sign in with openid

[Sign up](#)  
[Forgot your password?](#)  
[Didn't receive confirmation instructions?](#)

© 2012 - 2015 MicrobeDB.jp project.

# GenomeRefineを利用して見る

The screenshot shows a web browser window for 'Genome Refine' at the URL [genome.annotation.jp/genomerefine/](http://genome.annotation.jp/genomerefine/). The page has a light gray header bar with the title 'Genome Refine'. Below it is a dark gray navigation bar with the word 'About' in white. The main content area has a light gray background. At the top left of this area, there is a 'Menu' button with a dropdown arrow. A red arrow points from the bottom left towards this 'Menu' button. The menu itself is a white box containing the following items:

- プロジェクト一覧
- 新規作成
  - MiGAP
  - MeGAP
  - MeGAP一括

Below the menu, there is a section titled 'Functions and Features' with several bullet points:

- ユーザ・グループ認証機能  
OpenIDユーザ認証拡張によりグループ情報をもつ認証機能。参加グループのみでゲノムデータ共有され、公開・非公開が選択可能です。TogoAnnotationなどの連携サービスも同じアカウントで利用出来ます。
- ゲノムデータ入力・編集機能  
微生物ゲノムアノテーション実行パイプラインMiGAPの結果(fasta形式およびcsv形式)を入力します。
- メタデータ入力・編集機能  
GenBank形式ファイルおよび比較解析に必要な情報をユーザが入力します。Organismは入力必須項目です。
- データ形式変換機能  
Resource Description Framework (RDF)形式および標準的なGenBank, GFF3, FASTAで出力します。編集されたメタデータ、アノテーションが反映されます。
- 自動ゲノムアノテーション編集機能 (\*)  
NITE標準アノテーション辞書とテキストマイニング技術を用いた遺伝子産物名称のアノテーション高精度化ツールTogoAnnotatorを利用。自動によるアノテーションを高度化を実施します。
- ゲノムデータベース機能 (\*)  
CyanoBaseと同じデータベースシステムによるゲノムデータベースを提供。キーワード検索およびBlast検索が実行可能です。
- 手動ゲノムアノテーション編集機能 (\*)  
CyanoBaseと同じユーザインターフェースをもつデータベース上で手動アノテーション編集機能を提供。遺伝子ページから編集が可能で、データはTogoAnnotationに保存されます。

\* 開発中のため機能を制限しております。本機能を利用してみたい方は個別にご相談下さい。

Sementic WEB technology

# Genome Refine: Project List

Genome Refine

genome.annotation.jp/genomerefine/projects

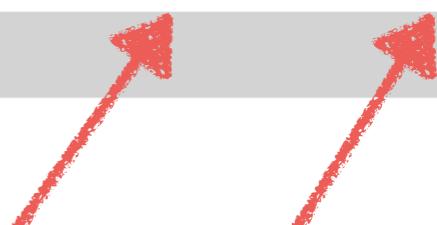
Sign out

## Project List

ID	Group	Type	Submission	Number of Sequence	Organism	Version	Created at	Updated at	Status	Genome Database	MicrobeDB.jp	Project
GR23	sample	MiGAP		1	Bacillus subtilis subsp. subtilis 168	9	2014-03-06	2014-04-02	failure	✓	✓	edit delete download
GR24	sample	MiGAP		1	Escherichia coli MG1655	1	2014-03-06	2014-03-06	processed	✓	✓	edit delete download

Action

- Project Create
  - MiGAP
  - MeGAP
  - MeGAP一括



# Genome Refine: Project List

Genome Refine

genome.annotation.jp/genomerefine/projects

Sign out

## Project List

ID	Group	Type	Submission	Number of Sequence	Organism	Version	Created at	Updated at	Status	Genome Database	MicrobeDB.jp	Project
GR23	sample	MiGAP		1	Bacillus subtilis subsp. subtilis 168	9	2014-03-06	2014-04-02	failure	✓	✓	edit delete download
GR24	sample	MiGAP		1	Escherichia coli MG1655	1	2014-03-06	2014-03-06	processed	✓	✓	edit delete download

Action

- Project Create
  - MiGAP
  - MeGAP
  - MeGAP一括



アノテーションDB

微生物統合DB MicrobeDB.jp

# GenomeRefineを利用して見る

The screenshot shows a web browser window for 'Genome Refine' at the URL [genome.annotation.jp/genomerefine/](http://genome.annotation.jp/genomerefine/). The page has a light gray header bar with the title 'Genome Refine'. Below it is a dark gray navigation bar with the word 'About' in white. The main content area has a light gray background. At the top left of this area, there is a 'Menu' button with a dropdown arrow. A red arrow points from the bottom left towards this 'Menu' button. The menu itself is a white box containing the following items:

- プロジェクト一覧
- 新規作成
  - MiGAP
  - MeGAP
  - MeGAP一括

Below the menu, there is a section titled 'Functions and Features' with several bullet points:

- ユーザ・グループ認証機能  
OpenIDユーザ認証拡張によりグループ情報をもつ認証機能。参加グループのみでゲノムデータ共有され、公開・非公開が選択可能です。TogoAnnotationなどの連携サービスも同じアカウントで利用出来ます。
- ゲノムデータ入力・編集機能  
微生物ゲノムアノテーション実行パイプラインMiGAPの結果(fasta形式およびcsv形式)を入力します。
- メタデータ入力・編集機能  
GenBank形式ファイルおよび比較解析に必要な情報をユーザが入力します。Organismは入力必須項目です。
- データ形式変換機能  
Resource Description Framework (RDF)形式および標準的なGenBank, GFF3, FASTAで出力します。編集されたメタデータ、アノテーションが反映されます。
- 自動ゲノムアノテーション編集機能 (\*)  
NITE標準アノテーション辞書とテキストマイニング技術を用いた遺伝子産物名称のアノテーション高精度化ツールTogoAnnotatorを利用。自動によるアノテーションを高度化を実施します。
- ゲノムデータベース機能 (\*)  
CyanoBaseと同じデータベースシステムによるゲノムデータベースを提供。キーワード検索およびBlast検索が実行可能です。
- 手動ゲノムアノテーション編集機能 (\*)  
CyanoBaseと同じユーザインターフェースをもつデータベース上で手動アノテーション編集機能を提供。遺伝子ページから編集が可能で、データはTogoAnnotationに保存されます。

\* 開発中のため機能を制限しております。本機能を利用してみたい方は個別にご相談下さい。

Sementic WEB technology

# MicrobeDB.jpで確認する

The screenshot shows a web browser window with the URL [microbedb.jp/MDB/migap/?group\\_id=241&project\\_id=GR24](http://microbedb.jp/MDB/migap/?group_id=241&project_id=GR24). The title bar says "MDB". The main content area displays two sections: "Migap Project Information" and "Migap Project Sequences", both showing "No items found". A red arrow points to the "Search" button in the top right corner of the "Migap Project Information" section.

Microbe DB.JP

Genome Refine 241/GR24 - Top MDB cyanobase/Synechocystis - s6803:slr0611

Search

Login is required to peruse this page. [Sign In](#)

Migap Project Information help

No items found

Migap Project Sequences help

No items found

# MicrobeDB.jpで確認する

The screenshot shows a web browser window for the MicrobeDB.jp website. The URL in the address bar is `microbedb.jp/MDB/migap/?group_id=241&project_id=GR24`. The page title is "MDB". The main content area displays "Migap Project Information" and "Migap Project Sequences". A red arrow points from the text "キーワード“dnaA”を入力" to the search bar, which contains "demo". The "Sign out" button is also highlighted with a red box.

**Migap Project Information**

Canonical Name	GR24
Division	BCT
Mol Type	Genomic DNA
Molecule Type	DNA
Topology	linear
Taxonomy Id	2
Classification	Bacteria
Visibility	private

**Migap Project Sequences**

Sequences	Label	Accession	Length
GR24-Chr1	sequence1.1	GR24-Chr1	4641652

# MicrobeDB.jpで確認する

The screenshot shows a web browser window for MicrobeDB.jp. The address bar indicates the URL is [microbedb.jp/MDB/search/?query=dnaA](http://microbedb.jp/MDB/search/?query=dnaA). The main content area displays a search result for the query "DNA". A sidebar on the left is titled "Disease" and shows "No items found". Below this is a table titled "MiGAP Gene" with the following data:

Locus Tag	Name	Synonym	Products	Notes	Project ID	Group Name
GR24_136450	GR24_136450		chromosomal replication initiator protein DnaA	gene_3645, P-loop containing Nucleoside <a href="#">...More</a>	<a href="http://genome.microbedb.jp/241/GR24 sample">http://genome.microbedb.jp/241/GR24 sample</a>	
GR24_124670	GR24_124670		DnaA regulatory inactivator Hda	DNA replication initiation factor; Provi... <a href="#">More</a>	<a href="http://genome.microbedb.jp/241/GR24 sample">http://genome.microbedb.jp/241/GR24 sample</a>	
GR24_130920	GR24_130920		DnaA initiator-associating protein DiaA	dimer interface [polypeptide binding], D... <a href="#">More</a>	<a href="http://genome.microbedb.jp/241/GR24 sample">http://genome.microbedb.jp/241/GR24 sample</a>	
GR23_10350	GR23_10350		DNA replication initiation control protein YabA	DNA replication initiation control protei... <a href="#">More</a>	<a href="http://genome.microbedb.jp/241/GR23 sample">http://genome.microbedb.jp/241/GR23 sample</a>	
GR23_10010	GR23_10010		chromosome replication initiator DnaA	Walker B motif, ATP binding site [chemic... <a href="#">More</a>	<a href="http://genome.microbedb.jp/241/GR23 sample">http://genome.microbedb.jp/241/GR23 sample</a>	
GR23_127280	GR23_127280		chromosomal replication initiator protein dnaA	hypothetical protein; Provisional, The A... <a href="#">More</a>	<a href="http://genome.microbedb.jp/241/GR23 sample">http://genome.microbedb.jp/241/GR23 sample</a>	

A red arrow points from the text "同一グループのデータが閲覧できる" to the "Group Name" column in the MiGAP Gene table.

同一グループのデータが  
閲覧できる

# Genome Refine:サービス間連携

