

MicrobeDB.jpポータル:

開発中のゲノム比較解析機能の紹介

MicrobeDB.jp

Home

Document

Analysis

e.g. hot spring, Enterococcus faecalis, nsbA

Search

guest

Comparative analysis:
Metagenomic samples
Taxon vs Taxon
Genome vs Genome
Environments (MEO vs MEO)

ゲノム比較解析（未公開）

MicrobeDB.jp

Integrating and representing genome, metagenome, taxonomy resources and the analysis datasets with Semantic Web Technologies.

Database statistics

Total number of Metagenomic samples (SRA/SRS):	173,359 samples
- with taxonomic analysis results:	60,551 samples
- with functional analysis results:	4,048 samples
Total number of Assembled Genomes (RefSeq/Genbank):	16,983 taxa
Total number of Strains (JCM/NBRC):	16,671 strains
Total number of Environmental terms in ontology (MEO):	2,381 terms

Show graph

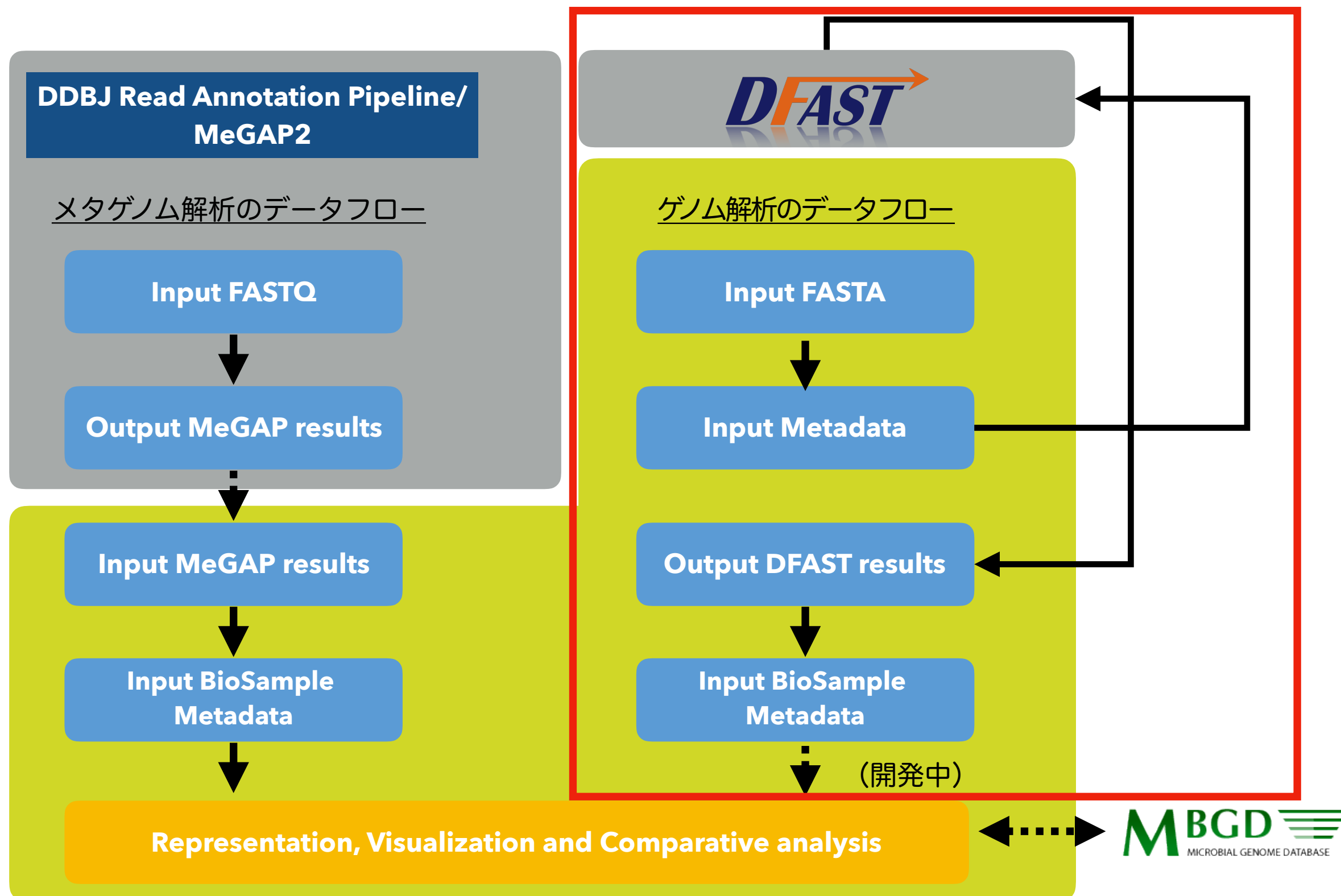
Features

Keyword Search

MicrobeDB.jp provides a keyword search function with a simple interface. The keyword search gives the user free-text access to the literal fields of all RDF/OWL resources on MicrobeDB.jp. See [document#sesource](#) for more information.

Representation and Visualization

MicrobeDB.jp解析パイプラインとの連携



MicrobeDB.jpポータル上からユーザ自身のメタゲノム・ゲノムデータ解析および比較解析を実現するため、それぞれメタゲノム解析パイプラインMeGAP、微生物ゲノムアノテーションパイプラインDFASTと連携させた。MeGAPについては非常に計算リソースを要求するため、国立遺伝学研究所スパコンで運用中のDDBJ Read Annotation Pipelineを介して実行後、解析結果が入力となる。

Genome analysis

DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication

Yasuhiro Tanizawa, Takatomo Fujisawa and Yasukazu Nakamura*

Center for Information Biology, National Institute of Genetics, Research Organization of Information and Systems,
1111 Yata, Mishima 411-8540, Japan

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on August 15, 2017; revised on October 27, 2017; editorial decision on October 30, 2017; accepted on November 6, 2017

ゲノム比較解析のためのデータアップロード機能

The screenshot displays the 'Genome Analysis' section of the beta.microbedb.jp website. The interface is divided into several sections for data submission:

- Submission Process:** The main heading for the upload section.
- Upload Sequences:** A form for uploading FASTA files. It includes fields for 'Title' (with a placeholder '20180306101105'), 'Group' (with a 'create new group' button), 'Data Privacy' (radio buttons for 'Private' and 'Public'), and 'Sequences' (a file upload area showing '0 files').
- Metadata:** A section for uploading Excel-formatted metadata files, including a 'Download Metadata' button.
- Download BioSample metadata template:** A section for downloading a template file, including a 'Download' button.
- Upload BioSample metadata:** A section for uploading BioSample metadata files, including a 'Metadata' field.

At the bottom, the URL https://beta.microbedb.jp/submissions/152/edit?entry_type=genome# is visible.

サインイン後にグループ
作成・管理、ユーザデー
タアップロードが可能

ゲノム配列およびDFAST
メタデータ (APIパラメー
ター情報)のアップロード

DDBJ BioSample登録
に準拠したBioSample
Packageと必須なメタデー
タ入力テンプレートのダ
ウンロード

DDBJ BioSample登録に
準拠したサンプルメタデー
タのアップロード

DFAST webサービス

https://dfast.nig.ac.jp

DFAST and DAGA are integrated genome annotation tools and resources. DFAST is an annotation platform for bacterial genomes. Its core annotation process is based on PROKKA and curated reference database tailored to specific organisms. It also generates DDBJ-compliant submission files for Mass Submission System (MSS) at DDBJ. DAGA is a genome archive that stores bacterial genomes obtained from DDBJ/ENA/GenBank and Sequence Read Archive (SRA). All the genomes deposited in DAGA are consistently annotated using DFAST. This website provides genome resources for *Lactic Acid Bacteria*.

DFAST

DDBJ Fast Annotation and Submission Tool

Upload your Genome, Annotate, and Submit to DDBJ.

Select reference database to use.

[Lactic Acid Bacteria](#) (v. 0.9)

β-versions

Cyanobacteria	Manually curated, based on CyanoBase . (v. 0.1)
E. coli	Manually curated. (v. 0.1)
RefSeq	For general use. Mainly constructed and automatically curated using protein sequences from 'Reference Genomes' in RefSeq.
Actinobacteria	Same as above. Subset for Actinobacteria.
Firmicutes	Same as above. Subset for Firmicutes.
Proteobacteria	Same as above. Subset for Proteobacteria.

You can see the Example of the annotation from [here](#).

Top page

DFAST is an annotation platform for bacterial genomes. Its core annotation process is based on PROKKA and curated reference database tailored to specific organisms. It also generates DDBJ-compliant submission files for Mass Submission System (MSS) at DDBJ.

Query File (Fasta format) **Job Title**

Mail Address

Specify metadata and parameters.

These data other than minimum contig length can be altered later. Reference Databases for genera other than *Lactobacillus* and *Pediococcus* are not fully supported.

Genus **Species** **Strain**

ex) *plantarum*, *delbrueckii* subsp. *bulgaricus*

Locus Tag Prefix **Minimum Contig Length**

Job submission

2. Input Metadata

Input metadata by filling the form to create the submission file. Please refer to the [instruction](#) for more information for each item. If you want to add items not shown in the form, you can add them manually after downloading the annotation file.

You can "Preview" the provided metadata. You can also import metadata from another DFAST job by providing the job ID in the text box below. By default, organism specific data, such as a species name or a strain name, is not imported. To enable this, check the "Override organism specific data".

登録ファイル作成に必要なメタデータをフォームに入力します。入力項目の詳細な説明や作成例については [アノテーションファイル作成概説](#) を参照してください。フォームにない項目はファイルダウンロード後に手動で修正を行ってください。メタデータを入力せずに空欄のままファイルを作成し、ダウンロード後に手動で入力することも可能です。

メタデータ入力後 "Preview" ボタンで確認ができます。また、ジョブ ID を入力して他のジョブからメタデータをコピーすることができます。デフォルトでは生物種名や菌株名などのデータはコピーされませんが、"Override organism specific data" を有効にした場合、これらのデータも上書きされます。

☐ : Override organism specific data.

Genus* **Species***

Strain* **Type Strain** **Culture Collection**

Locus Tag Prefix **BioProject*** **BioSample*** **Sequence Read Archive**

Submitter

Submitters*

DDBJ submission file editor

Archive Download Help

Edit a Feature

Feature 1 : LDL_00001

Product **Gene**

Note

☐ Highlight this Feature

Translation

MINGLVHWIAQNFPNIYNLWGTGDTGWGTSILQIIMTFWPSIFGGVLGLFF
GVILVLTEPGGILENKFWNFCDKLISILRAIPFIILLAFISPLTRLIVGTEIGDTAA
LVPLTLGIFPFYARQVQVALES LDPGKVEAAQSLGASNWDIIFDVYLKETRSEII
RVSTVSIISLIGLTAMAGAVGAGGLGTTAIQYGYRDFANDVTFLATVLVIMIFIV
QVVGDFLAKKLNHQHR

[Blast this sequence at NCBI.](#)

Seq. ID	Location	CDS	methionine ABC	metP	View
sequence01	complement(396..7				
sequence01	complement(794..1				
sequence01	complement(1556..				
sequence01	complement(3123..3836)	CDS	methionine ABC	metP	<input type="button" value="View"/>
			transporter permease		<input type="button" value="View"/>

Annotation editor

DFAST: DDBJ Fast Annotation and Submission Tool

V 0.1 (2016.1) Tanizawa et al., BMFH (2016)

軽量アノテーションパイプライン Prokka をベース

生物種群ごとにキュレーションされたDBを用意

メタデータ・アノテーション結果を編集するGUI

→ DDBJ compliant な登録ファイルの生成

Web APIによるジョブ投入・管理機能の実装

汎用的な参照DBを用意

独自アノテーションエンジンDFAST-coreの新規開発

Stand-alone ツールとしても公開 (GitHub: nigyta/dfast_core)

Tanizawa et al., Bioinformatics (2017)

V 1.0 (2017.8)

DFAST-core: 新バックグラウンドエンジン

スタンドアローンツールとしても利用可能

https://github.com/nigyta/dfast_core

Linux/Mac, Python 2.7/3.4- で動作

外部バイナリー同梱

カスタマイズ可能なワークフロー

高速かつリッチなアノテーション

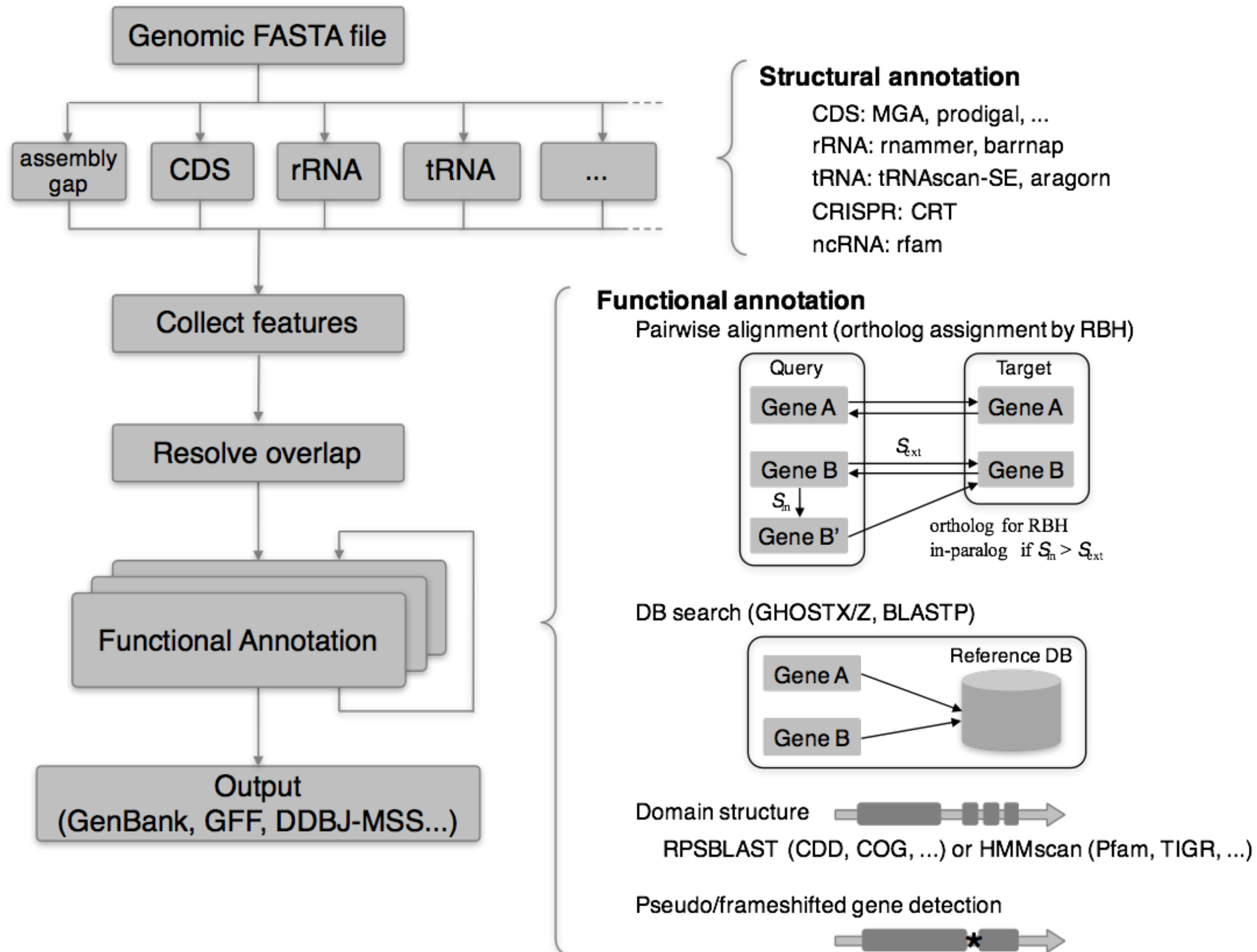
GHOSTX をデフォルトで利用

偽遺伝子・フレームシフトの検出

DDBJへの登録ファイル作成

DFAST-core アノテーションワークフロー

参照DB・外部ツール・パラメータ等を自由に設定可能



典型的なサイズのバクテリアゲノムを約5分でアノテーション

Prokka と同等の実行時間でより多くの遺伝子の機能を予測

参照DBのサイズはProkkaの20倍

BLASTPを使った場合の実行時間の1/10

セレノシステインを含む3遺伝子の検出に成功

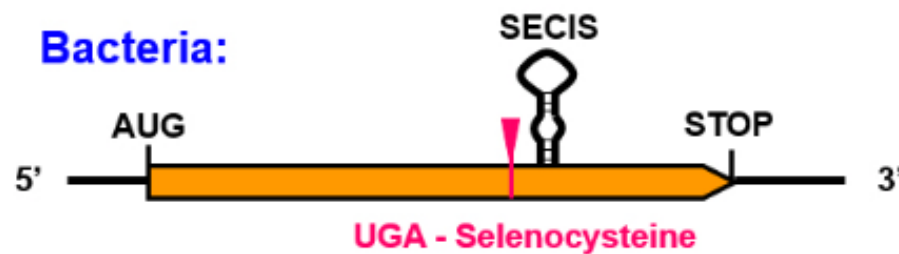
Escherichia coli O26 アノテーション結果

Data source / Annotation tool	INSDC	RefSeq (PGAP)	DFAST	Prokka	MiGAP
Total CDS	5795	6243	5740	5759	5721
<i>Pseudogene</i> frameshift or internal stop / partial	276	337 (250/87)	344 (158/186)	[30*]	-
<i>Selenoprotein</i>	3	1	3	-	-
<i>with COG number</i>	-	-	3965	-	4392
<i>Unknown function</i>	1203	1514	1347	2068	418
tRNA	101	101	105	105	100
rRNA	22	22	22	22	22
CRISPR arrays	-	2	2	2	-
Running time	-	-	3m27s	3m20s	4h43m

* Indicated as possible pseudo in the log file
(INSDC: GCA_000091005.1, RefSeq: GCF_000091005.1)

DFASTで対応しているアノテーション出力例

翻訳の例外



```
/inference="DESCRIPTION:similar to AA  
sequence:RefSeq:NP_310105.1"  
/transl_except=(pos:2026834..2026836,aa:Sec)  
/note="codon on position 196 is selenocysteine opal codon."  
/note="NP_310105.1 nitrate-inducible formate  
dehydrogenase-N alpha subunit (Escherichia coli O157:H7  
str. Sakai) [pid:99.8%, q_cov:100.0%, s_cov:100.0%, Eval:0.0e+00]"
```

偽遺伝子・フレームシフトの検出

Reference protein



Query



Reference protein



Query



```
/note="Partial hit; WP_003643223.1 gluconate permease  
(Lactobacillus plantarum WCFS1) [pid:71.3%, q_cov:100.0%,  
s_cov:44.8%, Eval:6.5e-78]"  
/note="frameshifted; deletion at around 14464"  
/product="hypothetical protein"
```

```
/inference="similar to AA sequence:RefSeq:WP_003548611.1"  
/inference="similar to AA sequence:MBGD:2015-01_default:16"  
/codon_start=1  
/product="AraC family transcriptional regulator"  
/transl_table=11  
/note="WP_003548611.1 AraC family transcriptional regulator  
(Lactobacillus acidophilus NCFM) [pid:36.0%, q_cov:99.7%,  
s_cov:99.4%, Eval:8.0e-54]"  
/note="MBGD: {gene_id: 'crn:CAR_RS11760', cluster_id: '16',  
gene_description: 'transcriptional regulator', version:  
'2015-01', tabid: 'default', pid: 40.2%, q_cov: 31.1%,  
s_cov: 29.8%, Eval: 1.1e-15}"
```

MBGD Cluster IDの
アサイン (開発中)

ゲノム比較解析インターフェース（イメージ）

MicrobeDB.jp

Home Document Analysis

e.g. hot spring, Enterococcus faecalis, psbA Search

guest

Search id or term...

Public/Private

- ☐ metagenome_public 173359
- ☐ metagenome_private 4

hasMetagenomeAnalysis

- ☐ taxonomy 60555
- ☐ function 4048

hasMEO (Text)

Search MEO terms ...

hasMEO: Component

Component for environment 36544

hasMEO: Env

Environment for microbes 166075

hasMEO: Position

Position toward environment 45406

hasMEO: State

State 96983

hasHostTaxonomy (Text)

Search HostTaxonomy...

hasHostTaxonomy

cellular organism 126540

uncultured 126540

Viruses 126540

pH

Temperature

Metagenomic samples 173363 results found in 849ms

Clear all filters

Previous 1 2 3 4 ... Next

10 Select All Deselect All

Select	MDB SampleID	msv:sampleTitle	msv:scientificName	msv:hasTaxonID	msv:hasBioProject	msv:hasBioProject	msv:hasBioProject
<input type="checkbox"/>	MDB000046		human oral metagenome	447426			
<input type="checkbox"/>	MDB000047		human oral metagenome	447426			
<input type="checkbox"/>	MDB000048		human oral metagenome	447426			
<input type="checkbox"/>	MDB000049		human oral metagenome	447426			
<input type="checkbox"/>	SRS101320	Human metagenome sample from G_DNA_Palatine Tonsils of a female participant in the dbGaP study "HMP Core Microbiome Sampling Protocol A (HMP-A)"	human metagenome	646099		SAMN00097841	2010-08-18T15:01:20Z
<input type="checkbox"/>	SRS101321	Human metagenome sample from G_DNA_Throat of a female participant in the dbGaP study "HMP Core Microbiome Sampling Protocol A (HMP-A)"	human metagenome	646099		SAMN00097842	2010-08-18T15:01:20Z
<input type="checkbox"/>	SRS101322	Human metagenome sample from G_DNA_Supragingival plaque of a female participant in the dbGaP study "HMP Core Microbiome Sampling Protocol A (HMP-A)"	human metagenome	646099		SAMN00097843	2010-08-18T15:01:20Z
<input type="checkbox"/>	SRS101323	Human metagenome sample from G_DNA_Subgingival plaque of a female participant in the dbGaP study "HMP Core Microbiome Sampling Protocol A (HMP-A)"	human metagenome	646099		SAMN00097844	2010-08-18T15:01:20Z
<input type="checkbox"/>	SRS101324	Human metagenome sample from G_DNA_L_Retroauricular crease of a female participant in the dbGaP study "HMP Core Microbiome Sampling Protocol A (HMP-A)"	human metagenome	646099		SAMN00097845	2010-08-18T15:01:20Z
<input type="checkbox"/>	SRS101325	Human metagenome sample from G_DNA_R_Retroauricular crease of a female participant in the dbGaP study "HMP Core Microbiome Sampling Protocol A (HMP-A)"	human metagenome	646099		SAMN00097846	2010-08-18T15:01:20Z

comparison analysis Compare 0 Please select items within range 2 - 100

MBGD MICROBIAL GENOME DATABASE

メタゲノム比較解析と同様のインターフェースで提供予定

公開データおよびユーザ自身のアップロードしたデータを選択

メタデータのファセット検索による絞り込み

MBGD連携した比較解析

課題

- ゲノムにおいては、MBGDオーソログ解析、メタゲノムにおいては、MeGAPメタゲノム解析（taxonomy/function）に大規模な計算リソースと時間を必要とするため、最新の公開データに関して解析データを含めたデータベースからの提供は困難である

今後の対応

- ゲノム、メタゲノムサンプルのメタデータについてはの更新を行い、最新の公開データを対象とした比較解析のためのMeGAP, DFASTの実行はユーザ自身に委ねる
- MBGD、MeGAPによる解析結果の反映についてはメジャーリリース時に対応